A model to explain why people act egoistically while feeling moral
# TABLE OF CONTENT

**INTRODUCTION** .......................................................................................................................... 3

**LITERATURE REVIEW** .................................................................................................................. 5

The traditional approach to social preferences and the moral wiggle room ........................................ 5
Self-signalling and moral sense .......................................................................................................... 7
Motivated reasoning and self-signalling .............................................................................................. 9
Separate equilibria ............................................................................................................................ 11

**THE MODEL** .................................................................................................................................. 12

A self-signalling model ......................................................................................................................... 12
Example: ............................................................................................................................................ 14

**FURTHER APPLICATIONS** .......................................................................................................... 17

Mental budgeting ............................................................................................................................... 17
Discrimination ..................................................................................................................................... 17
Blood donation .................................................................................................................................... 18

**CONCLUSION & FURTHER RESEARCH** ...................................................................................... 21

Comparison to the base model (Bodner & Prelec, 2002) .................................................................. 21
Conclusion .......................................................................................................................................... 21
Bibliography ....................................................................................................................................... 24
Appendix ............................................................................................................................................ 28
INTRODUCTION

Morality is a building block of any society in the world. It accompanies and supports most of our institutions and makes us accept and comply with rules imposed on us. For instance, even without legal punishment, one does not commit murder for no reason since it is hardly justifiable in terms of morality. In that sense, morality is not only a justification of the law but also the origin of the main principles that are defended by the most basic judiciary system. Of course, this does not mean that the judiciary system represents morality, but that its most simple principles are derived from a sense and intuition of what is moral, making morality the cement of our social institutions and a key element to study.

In economics, morality has often been modeled in two ways. In classical social preferences models, consequences of our actions are evaluated and then the morality of the actions are modelled by adding an extra term to the utility function, representing the effect of our actions on others (Ernst Fehr & Schmidt, 2003) - using a broad definition of morality that includes deviation from egoistic behaviours like fairness, honesty, reciprocity or altruism. Moral concerns can also be represented through self-signaling models, which explain these behaviours by the signal sent to the self when one acts morally. Individuals would not act morally because of the consequences on others, but simply because it feels good to be moral a person (Bodner & Prelec, 2002). Both types of models predict a tendency of pro-social behaviors.

Nevertheless, the mere existence of the judiciary system can also be considered as a symptom of the immoral tendencies societies have. But the fact that individuals engage in questionable behaviours is not surprising, even murder is normal from a sociological point of view (Durkheim, 1895), the society therefore had to establish extra incentives, or punishment to control the murder rate.

Immoral behaviours are often justified by exceptional circumstances “I had no choice”, because “everyone else was doing it” or “I was just following order”. The environment seems to have a significant role in the way we judge the morality of our actions and justify them. For example, when having to make a choice between buying a lottery ticket for a charity or to themselves individuals overestimate risk for the moral option to justify the egoistic choice. Similarly, they also tend to avoid donations request altogether as it seems that people are less comfortable declining a request than avoiding it. They seem to use diverted techniques in order to process
and judge their actions under a more favourable light. (Flynn & Lake, 2008; Lazear, Malmendier, & Weber, 2012).

These self-serving interpretations of reality or morality, aimed at the protection of the image of the self, lead to a situation where individuals “strongly believe they are just, virtuous, and moral; yet regard the average person as distinctly less so” (Tappin & McKay, 2016). Individuals can thus act immorally but find a way to justify these actions and get a moral sense nonetheless. These interpretations, that disconnect moral feelings from moral actions are hardly explained in the economics literature. Both social preferences and self-signalling theory rely on a direct causation: if someone acts morally, he feels moral or is happy to see the beneficial effects on others.

This thesis aims to fill this gap by explaining how individuals can exploit certain environments to get a sense of morality while acting egoistically. To achieve this, I will first review and try to bridge three different parts of the psychological and economic literature: motivated reasoning, social preferences and self-signaling. A model synthesising the literature review will then be constructed and several applications scenarios will be provided. Finally, I will discuss some of its hypotheses and compare the model to its source material: (Bodner & Prelec, 2002)
LITERATURE REVIEW

The traditional approach to social preferences and the moral wiggle room

Traditional economic models are built on the hypothesis that people are only pursuing material self-interest. The development of experimental economics in the 1980s with laboratory games such as the ultimatum game or the dictator game yielded results not consistent with the self-interest hypothesis and demonstrated that subjects also had social preferences: an interest in other subject’s well-being (Ernst Fehr & Schmidt, 2003).

Decades of research have since confirmed that individuals have such type of moral concerns and tend more generally to care about other people’s payoffs. These concerns were usually modelled by adding an extra term in utility functions that increased (decreased) utility if the concerns were (not) met (Bolton & Ockenfels, 2000; Charness & Rabin, 2002; E. Fehr & Schmidt, 1999).

These models are useful as they can capture the utility an individual gets from the causal effects of his or her prosocial behaviours. In the model developed below, such concerns will be formalised by the outcome utility. Many empirical evidence can be explained by such “outcome utilities”, but some accounts, especially where the decision-making context is complex cannot be understood by such a straightforward approach.

The most striking example comes from a study conducted by (Dana, Weber, & Kuang, 2007). Subjects were split into two groups, dictators (player X) and players (Y). Player X had to split the earning through a binary choice, option A and B. Choice B can arguably be considered as more moral as it features some sort of equity. As expected by classical social preferences empirical evidence, 74% of dictators chose option B.

![Figure 1: (Dana et al., 2007)](image)

Figure 1: (Dana et al., 2007)
In the other treatment, dictators were also asked to choose between two distributions. This time, a tweak was made to the experiment as subjects were not able to see player Y’s payoff. If they wanted to, the dictators could press a button before making their decision to reveal Y’s payoffs. Players had an equal chance to be in the situations 1 and 2 depicted in figure 2. Situation 1 had a similar payoff structure to the baseline experiment while the second made the egoistic and moral choice coincide. Results show that 63% of dictators chose the immoral option in the hidden information experiment when they were in situation 1. Only 56% decided to reveal player Y’s payoff.

![Figure 2: (Dana et al., 2007)](image)

This can hardly be explained by traditional social preferences models. If individuals had pure prosocial preferences, then all should have pressed the button to make sure the other player got a decent part of the endowment to ensure their concerns were met. Information should be weakly preferred over no information, hence “participants treat an action taken under willful ignorance as less indicative of an underlying egoistic motivation” (Gino, Norton, & Weber, 2016).

These results have since been replicated and consolidated in several experiments (Grossman & van der Weele, 2016; Larson & Capra, 2009; Matthey & Regner, 2011). The existence of a more opaque environment, where a direct causation between action and consequences is harder
to set, thus creates “a moral wiggle room” as called by (Dana et al., 2007), where individuals are more compelled by immoral choices.

Self-signalling and moral sense

The next question one must ask himself is how individuals can exploit this moral wiggle room. The answer provided in this paper relies on the assumption that people infer their quality and personal traits by observing their own behaviours. Some individuals would get extra utility, not because a moral action benefits other, but because the action reflects positively on themselves.

Self-signaling has been studied by psychologists, namely James-Lange’s famous theory of emotions. James-Lange hypothesised that individuals infer their emotional state by looking at their bodily response, for example shivering would be an indication of fear (Bodner & Prelec, 2002). But since information can also trigger an emotional response such as a hope or fear, the signals can also have an immediate impact on self-esteem (Bénabou, 2015; Bénabou & Tirole, 1999). In other words, information could have an immediate impact on utility through self-signalling.

(Quattrone & Tversky, 1984) showed evidence of such motivated self-signalling in an experiment, which took place in a medical facility to increase the credibility of the setup. They asked subjects to immerse their hand in freezing water for as long as possible. When debriefing, they explained in one group that tolerance to freezing water was positively linked to shorter lifespans while the other treatment received the opposite information. Subjects were then asked to repeat the tolerance test. Most subjects changed their tolerance to freezing water to fit the information that was given to them to “have a longer lifespan”. Giving people information about the signification of an action changed their effort to perform if it can benefit the image of their self.

Two main conclusions can be drawn from the experiment. First, individuals do interpret their actions and get (dis)utility from the diagnosis. Second, the value of this (dis)utility is important enough to significantly alter their effort. In many ways, the dilemma the subjects faced was similar to moral judgements. Resistance to freezing water can induce negative and positive emotions such as fear of death, happiness or pride. Moral emotions can be defined by a similar
array of valence ranging from shame and guilt to more positive emotions such as elevation and pride (Tangney, Stuewig, & Mashek, 2007).

These moral emotions are not only connected to the consequences on others but seem to be closely intertwined with moral judgement of the self. Making subjects recall immoral behaviours induced compensatory responses in subsequent moral activities, indicating that the remembrance of immoral behaviours had a negative signal impairing the image of the self that the subjects wanted to repair (Jordan, Mullen, & Murnighan, 2011). Likewise, when moral actions such as blood giving are stripped away from their potential positive signal by an extrinsic monetary motivation, individuals tend to be less motivated by the action leading to a crowding-out of donators, regardless of the unchanged causal effect on others. (Mellström & Johannesson, 2008).

Hence, if the traditional social preference literature cannot explain the account of the literature mentioned above, self-signalling, can provide an explanation of why individuals behave differently when the causal effects are the same. In a classical dictator game, individuals would not only give because of prosocial tendencies but also because of egoistic self-image concerns. In the model below, these concerned will be formalised by the diagnostic utility or the utility the individual gets from analysing his or her actions. For each action, a signal to the self is sent yielding a certain amount of utility. However, at this stage, self-signalling cannot interpret on its own why individuals act differently in presence or absence of a moral wiggle room.

The effect on self-esteem and the signal should presumably be the same, no matter how opaque the decision-making scenario is. If not, as already mentioned, individuals should weakly prefer information over no information and should, for example, push the button presented in Figure 2 of the (Dana et al., 2007) experiment before making their decision to maximise both their signal to the self and the causal effects of their actions. The next part will explain how individuals use an opaquer environment to justify immoral behaviours.
Motivated reasoning and self-signalling

A positive self-image is something that people care about and try to maintain through a lot of efforts (Kunda, 1990; Steele, 1988). If one could increase this self-image by logical and honest means like a fair outcome in a dictator game, it appears that individuals are also strongly biased in the interpretation of their actions.

Motivated reasoning is the result of the effect of motivation “through reliance on a biased set of cognitive processes: strategies for accessing, constructing and evaluating beliefs” (Kunda, 1990). In the self-signalling paradigm, this would mean that individuals’ diagnosis of their action will strongly be biased in a self-serving way to get a higher subsequent signal. An essential point here is that the ability to tweak the diagnosis of their action is not without limitation. Their environment constraints them to an action with a believable rationale (Kunda, 1990) if they are to expect any positive self-signalling.

For example, a standard dictator game with its extremely simple framework, would not provide an environment where individuals are able to use motivated reasoning as it would be hard to justify a completely egoistic choice. On the other hand, in presence of a moral wiggle room, the one to one relation between action and consequences is blurrier, allowing individuals to rely on a more biased and self-serving approach. Motivated reasoners could exploit moral wiggle rooms to construct a rationale justifying their actions and have a positive diagnosis of their actions. These rationales can be carried out in two main ways: self-serving judgment of morality and reality.

Self-serving judgement of morality occurs when an individual uses a definition of morality that specifically justifies his or her action. Morality being complex and hard to define, it is easy for individuals to transform or use a definition of morality that benefits the image of the self. Centuries of philosophical debates have produced many definitions of morality that can be used in a specific context. For example, one could use a Kantian definition of morality if it benefits himself or if the situation changes a utilitarian one. (Gino et al., 2016)

In controlled environments, (Messick & Sentis, 1979) showed that subjects fairness judgements were indeed influenced by the payment to the self. In other words, individuals were using judgements of what fairness is to maximise their own profit. Seeing alternative scenarios or
situations where their actions could be moral also seems to greatly affect people’s judgement of how moral their actions are. When individuals witnessed that a more beneficial counterfactual was possible, they were more likely to lie and misreport the outcome of a dice roll to enhance their gains (Shalvi, Dana, Handgraaf, & De Dreu, 2011). Similarly, if dictators also have the choice to take money from the other player in addition to just sharing their endowment, fair behaviours decrease significantly and most dictators simply keep the money that was given to them (Bardsley, 2008; List, 2007). Seeing that they could do worse, justified a sense morality.

Another way in which individuals can exploit a moral wiggle room and do motivated reasoning is by interpreting the reality in a self-serving way. These distorted interpretations can broadly be classified into two categories: wishful blindness and reality denial (Bénabou, 2015).

Wishful blindness occurs when an individual cherry picks his or her information to minimise the possible distress or disutility the information can carry. For example, patients with a Huntington disease, a degenerative disorder, often do not want to update or even get their diagnostic since it will most probably carry unwelcome news (Dorsey et al., 2013). Similar behaviours have been observed by (Dana, Cain, & Dawes, 2006) in dictator game experiments. In this set-up, participants were allowed to refuse to participate in the game if they were willing to give up 1€ of their initial 10€. Results show that about one-third of the subjects chose to opt out and keep 9€. Self-interested subjects would have rationally chosen the 10€, and if they really had moral concerns, they could have given the 1€ to the other player. These subjects preferred to give up 1€, in order not to be confronted by information about their morality.

Finally, reality denial happens when someone already got the unwelcome information but interpret it in a distorted way to avoid distress or discomfort. Reality denial behaviour can also lead to a certain “unethical amnesia”, where individuals obfuscate their misdeeds and unpleasant events to protect their image of the self (Anderson & Hanslmayr, 2014; Kouchaki & Gino, 2016). Uncertainty, complexity and situations where a one to one mapping between cause and effect is difficult to establish or can be avoided, like in a presence of a moral wiggle room, could possibly enable such distorted and self-serving interpretation of reality.
Knowing that immoral actions can have a detrimental effect on the image of the self, individuals can alleviate the weight of a negative self-signal or transform it into a positive one by convincing themselves that from a certain moral point of view, or by interpreting the reality differently. In a moral wiggle room, such reasoning would be facilitated by the less transparent decision-making structure as more room is left to interpret the action in a favourable way. In the next part, this mechanism will be formalized by the Bayesian updating the agent will undergo when deciding; or more precisely by the probability of getting a signal with and without a wiggle room.

**Separate equilibria**

The situations in which individuals must choose between prosocial and egoistic choices in the experiment described above may vary, but one key characteristic remains: some individuals will exploit moral wiggle rooms when presented, while others will not. For example, in (Dana et al., 2007), across all the treatment that implemented some sort of wiggle room, roughly 35% of individuals chose the fairest option. These individuals were not sensible to the temptation of the egoistic option and would choose the moral option regardless of the environment.

Since the causal effects of the actions remain the same with and without the moral wiggle room, the existence of separate equilibria imply that some individuals give more importance in term of utility to the self-signalling than others. This can be interpreted as the personal preferences and relative weight in the utility of the causal effects of the action compared to the self-image. In the model, the level of morality will make this distinction between individuals. For example, compared to less moral agents, agents with high levels of morality may care more about the real consequences of their choices and may derive less signalling utility from the same choices.
THE MODEL

A self-signalling model

The model is inspired by the self-signalling model developed by (Bodner & Prelec, 2002)\(^1\) where an individual tends to maximise a utility function that combines two psychological entities: the doer and the interpreter. The individual can choose an outcome \(x\) from a set of discrete options: \(A \in \mathbb{R}\) in an environment \(E \in \{0; 1\}\) such that \(A \perp E\). The Utility is composed of two elements: the outcome utility and the diagnostic utility.

\[
U(x, A, E, \theta) = v(x, \theta) + E[u(S_\theta)|x, A, E] \quad (1)
\]

The outcome utility \(v(x, \theta)\) is the classical part of the utility function, where an individual gets a certain amount of utility by choosing a certain option. It is also dependent on \(\theta\), an index representing a certain principle, in our case morality. The higher \(\theta\) the higher the moral standard of the individual. It is important to note that “The gut knows, but the mind does not.” (Bodner & Prelec, 2002), meaning that the individual is not able to take an expectation of \(v(x, \theta)\) as if \(\theta\) was a normal unknown parameter. \(A\) and \(E\) do not influence the outcome utility, \(v(x, \theta)\) represents the utility the individual gets from inferring the causal effects of his or her choice \(x\) regardless of the game structure.

The second part of the utility function is the diagnostic utility, \(E[u(S_\theta)|x, A, E]\), which is an expected utility that the signal \(S_\theta\) will yield, given the action and the choice set and the environment \(E\).

\(S_\theta = S(\theta)\) represents the signal the agent receives, it is a monotonous decreasing function of morality \(\theta\). This property is inspired by the public economics literature and the impure altruism model (Andreoni, 1989), where impure altruist use “apparent “charitable” behaviour [that] can also be motivated by a desire […] to receive social acclaim” (Becker, 1974). In the impure altruism model, next to the egoistic consumption of the endowment, donations to public goods enters the utility function twice, once as public good and once as a private good.

\(^1\) As a reminder, Prelec’s self-signalling model is defined the following way: \(U(x, A, \theta) = u(x, \theta) + \sum_\theta f(\theta|x, A)v(\theta)\)
In our case, the moral investment of an action can be divided into three parts. The first two are represented by the outcome utility, where the agent’s utility is impacted by the causal effects of his or her actions on him or her (the egoistic part) and on others (the altruistic component). The private part of the public donation would then be the diagnostic utility, representing the expected utility of the signal to him or herself. $\theta$ can be interpreted as the relative preference of the public good part compared to the private good part. An individual with low $\theta$ would only be driven to invest in a public good because of egoistic reasons: the signal; while a high $\theta$ would mean that $S_{\theta} = 0$ the individual would then just take the causal effects of the actions into consideration

$$E[u(S_{\theta})|x, A, E] = \int u(S_{\theta})p(S_{\theta} | x, A, E)dS_{\theta} \quad (2)$$

$$p(S_{\theta}|x, A, E) = \frac{p(x|A, E, S_{\theta})p(S_{\theta}|E)}{p(x|A, E)} \quad (3)$$

The updated probability depends on several factors. First, $p(x|A, E, S_{\theta})$ represents the probability of choosing $x$, given the choice set $A$, the environment $E$ and the signal $S_{\theta}$ that it would send. Similarly, $p(x|A, E)$ represents the probability of choosing $x$ given the choice set and the environment. We will consider that $E$ only affects the probability of choosing $x$ through the signalling such that $p(x|A, E) = p(x|A)$

Finally, $p(S_{\theta}|E)$ express the probability of getting a signal given the environment. In this work, the environment will represent the moral wiggle room.

Looking closer at the expected utility, we can see that the utility an individual can receive by diagnosing his action is dependent on different main variables. First, $\theta$, his morality, a high morality yields lower signals. Then the action $x$, the environment $E$ and the choice set $A$. The choice set and the environment are particularly important as it captures two elements that the outcome utility misses:

- How the agent considers the choice set before his or her action. For example, one might feel moral by choosing an option that gives half of his or her payoffs to another person.

---

2 See appendix 1
But another situation where this option is the less altruistic of them all, the individual would most likely not get as much utility in the second case.

- The presence or not of a moral wiggle room, that affects the moral signalling, and the choice, through \( p(S_\theta|E) \) ceteris paribus so that \( p(S_\theta|E = 0) \leq p(S_\theta|E = 1) \)

However, contrary to (Bodner & Prelec, 2002) \( S_\theta \) does not represent a value that an individual possesses, but a target of morality he or she wants to attain. Thus, the individual does not get his or utility thanks to the updating of the probability of having or not a certain quality, but because they will meet their “target of morality”. Recent evidence mentioned in the literature review that individuals do not recall (Kouchaki & Gino, 2016) their unethical behaviours on the long term – and therefore do not update their quality - can be interpreted in favour of this type of signalling.

Example:

Let us use classical dictator game where dictators must share 10€ between themselves (player \( i \)) and another player. In this case, \( \nu(x, \theta) \) can be modelized by a Fehr-Schmidt utility function, where \( x_i \) represents the dictator’s payoff and \( x_j \) the other player’s payoff. To continue the example, let’s assume that \( x_i \geq x_j \). The dictator’s utility function can be represented by:

\[
U(x_i) = x_i - \theta (x_i - x_j)
\]

We therefore have the following optimisation problem:

\[
U(x_i) = x_i(1 - 2\theta) + 10\theta \\
\text{s.t } 5 \leq x_i \leq 10
\]

And the solution are “the extremes” of the distribution of payoffs, namely

\[
x_i = \begin{cases} 
10 & \text{if } \theta < 0.5 \\
5 & \text{if } \theta > 0.5
\end{cases}
\]

In other words, dictators with \( \theta \) strictly inferior to 0.5, will prefer to go for the more egoistic option taking the entire endowment, while the others with \( \theta > 0.5 \), would still choose the fair option and give half of the of the endowment.
Prelec’s self-signalling model allows for a more nuanced approach that is more realistic when taking experimental evidence into account, namely that dictators tend to give a little bit less than 30% of their endowment to the other player (Engel, 2011). Individuals with $\theta < 0.5$, would give more than 0 as their diagnostic utility would then be positive and create a counterweight to compensate the loss of outcome utility. Dictator with $\theta > 0.5$, would still choose the fairest option as it would both maximise their outcome utility and diagnostic utility.

Let us now consider our model. We can describe the choice set $A \in \mathbb{R}$ let set the environment value as $E = 1$ in presence of a wiggle room and 0 otherwise. The simplicity of a standard dictator game can be interpreted as an environment with $E = 0$.

To continue the illustration, let us continue the example with 3 different actions such as $A = \{5; 7.5; 10\}$, two levels of morality $\theta = \{0; 1\}$, two types of signals, High and Low $S_\theta = \{H; L\}$ such that $u(S_\theta) = 5$ when $S_\theta = H$ and 0 otherwise. With the probability of getting a high or low signal presented in table 1, we can compute total utility, using the same optimisation constraint as in (3) with the corresponding $\theta$ for $E=1$ and $E=0$.

<table>
<thead>
<tr>
<th>$X$</th>
<th>$p(S_\theta = L)$</th>
<th>$p(S_\theta = H)$</th>
<th>$E[u(S_\theta)]$</th>
<th>$p(S_\theta = L)$</th>
<th>$p(S_\theta = H)$</th>
<th>$E[u(S_\theta)]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0.2</td>
<td>0.8</td>
<td>4</td>
</tr>
<tr>
<td>7.5</td>
<td>0.2</td>
<td>0.8</td>
<td>4</td>
<td>0.8</td>
<td>0.2</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0.8</td>
<td>0.2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: Probability per type of signal and expected utility per level of morality when $E=0$

In the absence of a wiggle room ($E=0$), agents cannot exploit reasons to justify their actions, their diagnostic utility decreases sharply when their choice is more egoistic. High morality individuals tend to have lower diagnostic utility as the probability of getting $S_\theta = H$, is lower for all $x$.

Table 2: Utility per level of morality when $E=0$

| $X$ | $v(x, \theta)$ | $E[u(S_\theta) | x, A, E]$ | $U(x, A, E, \theta)$ | $v(x, \theta)$ | $E[u(S_\theta) | x, A, E]$ | $U(x, A, E, \theta)$ |
|-----|----------------|-----------------------------|----------------------|----------------|-----------------------------|----------------------|
| 5   | 5              | 5                           | 10                   | 5              | 4                           | 9                    |
| 7.5 | 7.5            | 4                           | 11.5                 | 2.5            | 1                           | 3.5                  |
| 10  | 10             | 1                           | 11                   | 0              | 0                           | 0                    |
Consequently, high morality individuals will choose $x=5$ as it both maximises outcome and diagnostic utility individually. Low morality individuals will choose $x=7.5$, as it maximises total utility, even though it is a compromise between the outcome and diagnostic utility. This behaviour would be in accordance with (Bodner & Prelec, 2002), and respect empirical observations of dictator games behaviours. Additionally, this also allows for a more accurate description than (E. Fehr & Schmidt, 1999) as individuals have $x=7.5$ as an equilibrium choice.

Table 3: Probability per type of signal and expected utility per level of morality when $E=1$

| X    | $p(S_\theta = L)$ | $p(S_\theta = H)$ | $E[u(S_\theta)]$ | $p(S_\theta = L)$ | $p(S_\theta = H)$ | $E[u(S_\theta)]$
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0.1</td>
<td>0.9</td>
<td>4.5</td>
</tr>
<tr>
<td>7.5</td>
<td>0.1</td>
<td>0.9</td>
<td>4.5</td>
<td>0.7</td>
<td>0.3</td>
<td>1.5</td>
</tr>
<tr>
<td>10</td>
<td>0.4</td>
<td>0.6</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4: Utility per level of morality when $E=1$

| X    | $v(x, \theta)$ | $E[u(S_\theta)\mid x, A, E]$ | $U(x, A, E, \theta)$ | $v(x, \theta)$ | $E[u(S_\theta)\mid x, A, E]$ | $U(x, A, E, \theta)$
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>5</td>
<td>4.5</td>
<td>9.5</td>
</tr>
<tr>
<td>7.5</td>
<td>7.5</td>
<td>4.5</td>
<td>12</td>
<td>2.5</td>
<td>1.5</td>
<td>6.5</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>3</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

In presence of a wiggle room, the probability of having high signal increases as individuals are able to construct self-serving judgement of morality or reality. In other words, $p(S_\theta = L\mid E = 0) < p(S_\theta = L\mid E = 1)$ and the expected utility increases for all level of morality.

In this environment, patterns of behaviours change as low morality individuals now have a stronger diagnostic utility for low moral choices. The optimal choice for them is now the most egoistic option, while high morality individuals still maximise their utility by choosing the fair option.
FURTHER APPLICATIONS

Discrimination

Even though most societies condemn discriminatory behaviours publicly, these behaviours persist. The justification for these behaviours seems to play an important role: if people have another justification for discriminatory behaviours than pure prejudice, then the action would not appear discriminatory to them anymore. (Snyder, Kleck, Strenta, & Mentzer, 1979), illustrated the phenomenon in an interesting experiment where individuals had to choose between two movies. Each movie was projected in an identical and separate room, and each room included another spectator. In one of the room this spectator was acting as a disabled person. In the control group, the movies projected in the two rooms were identical. In the treatment group, each room had a different movie. The movie and the actor playing the disabled person were of course swaped, to control for these factors. Results showed that when the movies were identical, 58% sat with the disabled person while only 17% did when the movies were different. Subjects significantly increased their avoidance of the disabled person when the movies were different even when controlling for the movie, showing that they used the movie as an excuse to avoid the disabled person.

In the model, the treatment where different movies were projected can be considered as a change in the environment allowing a positive signal to the self. If $\theta$ represents prejudice towards disabled persons, then the outcome utility of a prejudiced agent would be significantly lower when a disabled person in the room. On the other hand, avoiding being with the disabled person without any “external” excuse or moral wiggle room, would also send a negative signal to the self. In this situation, some individuals would want to protect the image of the self and go in the room with the disabled person. In treatment 2, the individuals with high signalling would have an external excuse: they prefer the other movie. They could maximise their outcome utility by avoiding the disabled individuals while having an excuse in order not to get the disutility from the signalling.
Blood donation

Already mentioned in the literature review, the crowding-out of blood donators, when an incentive system is introduced, is also an interesting situation where the model can be applied. In the first situation, when no incentive system is introduced, individuals have two reasons to give their blood: first to help other people and second, to signal to themselves that they are doing something good. Interestingly, the introduction of the incentive systems seems to cancel the diagnostic utility they get from doing something good. This case, in opposition to the other situations studied in this thesis where individuals try to justify immoral behaviours by using their environment, depicts a situation where the signal is the driver of prosocial behaviours.

In the model, let us broadly define \( \theta \) as the desire to act prosocially. In the case without incentives, the donators will get utility from the outcome utility, because their action has a positive consequence on others and from the diagnostic utility: they feel good because the action reflects positively on themselves. On the other hand, when the incentives system is put in place, the diagnostic utility disappears, as the donator cannot reasonably convince himself he is only doing that out of pure goodness.

One might ask whether this disappearance comes from the change of the choice set or the environment. Even though the choice set does indeed change I argue that the change comes from the environment. The probability of getting a signal in the diagnostic utility depends on the utility of choosing one outcome over all the other choices. In the discussed case, this would mean, choosing to donate blood over not donating in the first situation, while choosing to donate blood for money over not donating in the second. From that point of view and with a given choice set, it is harder to discern which choice is the more prosocial. At the same time, regardless of the choice set, the source of the motivation is harder to discern in the second case as it could come from the extrinsic monetary motivation or the prosocial attitude of the individual. It is thus the environment, that is dependent on and arises from the choice and not the comparison to the choice set that creates an ambiguity inducing a weakened diagnostic utility.

The crowding out of intrinsic motivation, or diagnostic utility, also shows us how determinant self-signalling can be to induce prosocial behaviours. Even though, they did not consider the
effect of environment, this self-signalling interpretation of the crowding-out of motivation has also been recently supported by experimental evidence (Dubé, Luo, & Fang, 2017)

Mental budgeting

Mental accounting has been one of the most prominent developments in behavioural economics. Instead of having a global view of their income, people tend to create budgets for specific types of items, like groceries, restaurant, sport, cinema etc… The introduction, of what Thaler calls the purchase evaluation device in a utility function explained why individuals would have such behaviours. Individuals would then only buy a good z if the utility of the purchase evaluation device divided by the price is superior or equal to k, a constant representing the budget constraint for that good. This added both psychological realism and explanatory power when compared to standard consumer theory (Thaler, 1985). If the budget constraint was not met, that would mean that individuals suffer from self-control issues (Thaler & Shefrin, 1981).

The model presented here could also provide a complementary explanation to the phenomenon. The variable x can represent the quantity of the good multiplied by the price while θ represent the financial responsibility of the consumer. The outcome utility would then model the “material consequences” of the action, notably the consumption and the consequences on the budget.

The diagnostic utility could be considered as the expected utility of the signal representing monetary responsibility. For example, with E=1 when the individual is in his budget and E=0 for any out of the budget expense.

If the self-signalling approach would not bring a substantial advantage in terms of power or scope when compared to Thaler’s model, it would, however, bring more realism in the process it describes. By relating the purchase to a value the individual wants to attain, it would explain why a good i has a certain budget k at a time t. For example, while purchasing foie gras is not acceptable on a daily basis (E=0) whatever the price, it is certainly enjoyable when having a

---

3 The diagnostic utility could also take concerns such as the comparison to a reference price into account in order to keep Thaler’s insights. One would not be financially responsible if the price was too far away from a certain reference price for example.
special occasion such as Christmas (E=1). The latter occasion would not impair the value of financial responsibility, as it is normal from a cultural point of view, even for a responsible person, to get more expensive food for a special occasion.

Moreover, if the traditional mental accounting approach can explain why individuals sometimes go over budget, the model does not explain why subsequent guilt can be felt after compulsive purchases (Piron, 1991). Alternatively, in the model presented here, the effect of framing on self-control could also be modelled by the diagnostic utility. In this case, individuals would use the framed environment in an analogous way to the moral wiggle room to justify their actions. This approach could also possibly explain the guilt felt by certain people after compulsive purchase. Their expected utility of the signal that has been “tweaked” favourably by the wiggle room would have been different from the actual signal they got after diagnosing their action later.
CONCLUSION & FURTHER RESEARCH

Comparison to the base model (Bodner & Prelec, 2002)

This part will follow the methodology developed (Rabin, 2013) to assess new models, namely by analysing how the changes affect the power - or the sharpness of its prediction - compared to the baseline model and scope, that is to say, the number of situations it applies to.

By interpreting (Bodner & Prelec, 2002) stricto sensu, behaviours would be the same in the presence or absence of a wiggle room since the environment does not have an effect on utility. The current model would have more power as it explains the different behaviours when individuals are in the presence of a moral wiggle room.

If we take a broader interpretation of Prelec’s self-signalling model and consider that it could encompass the moral wiggle room, the dictator will have to maximise both terms of the utility function. With: 

\[ U(x, A, \theta) = u(x, \theta) + \sum_{\theta} f(\theta|x, A)\nu(\theta), \] 

if \( \nu(\theta) \) represents the utility an individual gets from being moral, then with \( \frac{dv(\theta)}{d\theta} \geq 0 \) and \( \frac{d^2v(\theta)}{d\theta^2} \leq 0 \), the more moral the individual is, the stronger the diagnostic utility. By considering that \( A \), encompasses the wiggle room, and that \( \sum_{\theta} f(\theta|x, A)\nu(\theta) \) increases in presence of a wiggle room, then paradoxically, the individuals who would be the most tempted to cheat and use the wiggle room would be those with high morality as their diagnostic utility would be the one with the most weight.

The model might still be able to represent the situations, but it would lose the psychological insights, making it a paramorphic model, that “describes the empirical phenomena of interest correctly, but the processes underlying the empirical phenomena are not matched by processes in the model” (Wakker, 2010). In other words, the change in the type of signalling allows to have a more realistic approach to the phenomenon and keeps the homeomorphism – or the accurate description of the underlying processes (Wakker, 2010) - of the source material when applied to this particular situation.

Conclusion

If traditional social preferences and self-signalling models can describe simple and straightforward decision-making scenarios, it appears that these models are not robust when the complexity of the environment is increased.
Starting from self-signalling decision-making model, this thesis has developed a model capable of explaining how individuals use their environment to justify immoral behaviours. The division of the utility into two categories, the outcome utility, based on causal effects and diagnostic utility based on the effect on the image of the self, seems to bring substantial benefits when compared to traditional social preferences models. Combined with motivated reasoning, the diagnostic utility can explain how individuals tend to use self-serving interpretations of morality and reality when the environment becomes opaque and why they do not when the environment is simple.

Of course, this work is not without limitations. Several assumptions have been made in the modelisation, the most important being that individuals follow Bayesian updating logic when making decision. This choice is a strong assumption that would need to be investigated, as it implies that the agent follows a strict logic in its updating. However, it seems to bear some psychological insights as individuals also tend to regret and feel shame after a decision was made, as if they were not completely able to forecast what the signal of their action will be.

Moreover, within the Bayesian updating, I considered that \( p(x|A,E) = p(x|A) \) such as the environment only affects the probability of choosing x through the signalling. This hypothesis is vital since \( p(x|A,E) \) is in the numerator. If x is an immoral choice \( p(S_θ|x,A,E) \) should increase within a moral wiggle room. Then, if \( p(x|A,E) \) also increases because of other reasons, then \( p(S_θ|x,A,E) \) would decrease, making the immoral option less attractive.

One could question this hypothesis, since the more complex environment that often accompanies moral wiggle rooms could also increase the probability of not understanding everything, and particularly one choice, for example, x. The agent would then consider it as less appealing, decreasing the probability of choosing x. However, in the experiment and decision-making scenarios described, such concerns should apply to all choices and there is no reason to think, that the immoral, or moral choices should be more affected by the environment. In other words, if the environment has a similar effect on the probability of choosing x, y or z, then \( p(x|A,E) \) should remain unchanged when E changes.
If this thesis provides several application scenarios to the model, such as donation behaviours, discriminations and mental budgeting, numerous other applications can be found, since any decision-making process where another personal dimension is involved could be modelled by the diagnostic utility. Further research on moral wiggle rooms is also necessary to study more precisely how the can enter the model. For example, some questions such as how does a set of choices affects an environment and vice versa could be useful to understand the relation between environment, utility and image of the self.
Bibliography


Appendix

\[
p(S_\theta | x, A, E) = \frac{p_0(S_\theta) \cdot p(x, A, E | S_\theta)}{p(x, A, E)} = \frac{p_0(S_\theta) \cdot p(x | A, E, S_\theta) \cdot p(A, E | S_\theta)}{p(x, A, E)} =
\]

\[
\frac{p_0(S_\theta) \cdot p(x | A, E, S_\theta) \cdot p(A, E | S_\theta) \cdot p(E | S_\theta)}{p(x, A, E) \cdot p(A | E) \cdot p(E | S_\theta)} = \frac{p(x | A, E, S_\theta) \cdot p(A | S_\theta) \cdot p(S_\theta | E)}{p(x, A, E) \cdot p(A)} = \frac{p(x | A, E, S_\theta) \cdot p(S_\theta | E)}{p(x, A, E)} \quad \text{since}
\]

the signal only depends on the environment by construction.