Deal-Making and Soft Commitment - a Behavioural Choice Model

Julie Sandilands

A dissertation submitted to Auckland University of Technology in partial fulfilment of the requirements for the degree of Master of Business

Lodged for examination in 2020

Business School

Economics Department

"Economists will and should be ignored if we continue to insist that … constantly trading stocks or accumulating consumer debt or becoming a heroin addict must be optimal for the people doing these things merely because they have chosen to do it."

Ted O'Donoghue & Matthew Rabin, 2003

Abstract

Individuals regularly indulge in internal 'deal-making' when justifying decisions they know to be bad. In the case of unhealthy foods, the justification often takes the form: "It's ok if I overindulge today, because I will eat well starting from tomorrow." This (soft) commitment to restrict future eating is usually reneged on once the future arrives, and the individual continues to overconsume unhealthy foods.

Standard choice theory states that individuals always make optimal choices, and therefore does not allow for the existence of these internal deals. Commonly used behavioural models account for the difference between optimal and observed choices in different ways, for example through 'present bias', limited attention, or through 'cue-triggered' decisions. To date, no model has incorporated internal deals as a cause of sub-optimal consumption choices.

This dissertation provides a critical synthesis of the literature on behavioural economic modelling, with a focus on the consequences of unhealthy food choices, and presents a model of internal deal-making as an alternative to the existing rational-choice and behavioural models of utility maximisation. The individual uses a soft commitment with the dual aims of satisfying cravings and achieving good health outcomes *on average*, but the inability to stick to this commitment lowers life-time utility. Choices and utility differ between naïve and sophisticated consumers.

The welfare impacts, key assumptions, and policy implications of the model are analysed and compared to existing models.

Table of Contents

List of Figures

List of Tables

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Acknowledgements

Thanks to Matthew Ryan and Peer Skov for supervising this dissertation.

Chapter 1: Introduction

Standard economics (which assumes 'rational' decision making) concludes that the individual is the best decision maker, as the individual is best placed to know their own preferences and the choices that optimise their well-being.

However, there is little evidence that people *are* rational in their decision making (Thaler and Sunstein, 2003), and our choices can often be improved upon. It's not just our mothers (or some other disapproving third party) who doubt our decisions. We ourselves often regret the choices we make; something that rational choice theory denies.

Studies continue to identify ways that humans deviate from economically-defined 'rationality'[1]. In response, behavioural economists have developed models that reflect our decision-making biases, and that measure the differences in welfare between what we choose, and what actually makes us better off.

Many of these models identify a 'true' utility function, which represents the utility we would choose to optimise if we were free of behavioural biases, and a 'decision' utility function which rationalises our actual choices made in the context of our biases.

Taken together, these models cover a range of biases, including present bias, addiction (the tendency for past consumption to influence present consumption), limited attention, and the influence of impulsive states which override our normal rational decision-making processes. Temptation, projection bias, and focusing have also been applied as the key cause of sub-optimisation in models of behavioural decision-making[2].

The differences in these models and the assumptions that underly them are important to understand and verify empirically, because they call for radically different policy solutions to combat the biases and return the individual to their optimal consumption levels.

Models of present bias and of focusing call for a tax on unhealthy items to overcome the effects of the bias and ensure people consume at the optimal trade-off between taste enjoyment and health harms[3]. In the temptation model proposed by Gul and Pesendorfer (2001) however, the correct approach to improve welfare is to remove tempting choices (by banning the sale of ice cream, for example) in order to improve individual health without losing utility due to temptation. The existence of environmental triggers in the Bernheim and Rangel (2004) model implies that removing environmental cues is a way to improve welfare, for example by banning advertising. Taxing unhealthy goods in this model is ineffective, as purchases while in the 'hot' state don't respond to price. In fact, taxes will make the individual worse off, as they will become poorer without changing their unhealthy behaviour.

So, should we tax unhealthy items? Ban them? Ban advertising for them? Farhi and Gabaix (2020) conclude that when people exhibit present bias towards unhealthy goods, we should tax the items when consumption is highly responsive to price (elastic), but look for other means of altering behaviour when consumption is inelastic. But what if present bias isn't the driving force in the decision-making process?

These aren't merely hypothetical questions. The World Health Organisation ("Obesity and Overweight," 2020) states that, in 2016, 39% of adults worldwide were overweight, and 13% were obese. According to the Ministry of Health, almost a third of New Zealand adults are

[1] See Frederick et al. (2002) and Bernheim & Rangel (2007b), for an overview of several key studies
[2] An overview of these models is presented on page 12.
[3] See, for example, O'Donoghue and Rabin (2006), Farhi and Gabaix (2020), Koszegi and Szeidl (2013)

obese (Ministry of Health, "Obesity Statistics," 2019). Given that obesity makes a person 7.2 times more likely to have type 2 diabetes (Abdullah et al., 2010, as cited in Harkanen et al., 2014) and 1.8 times more likely to have heart disease (Bogers et al., 2007, as cited in Harkanen et al., 2014), as well as more likely to suffer from a range of cancers and other health concerns, the obesity epidemic has serious consequences for human welfare.

Some of these individuals may be overweight or obese because of non-diet and lifestyle factors (i.e. genetics, medications, etc). In addition, some individuals may optimally choose diets that lead to poor health outcomes. These individuals love unhealthy food so much that the consequences of diabetes, heart disease and early death are an acceptable price to pay in exchange for all that delicious junk food. Even as their waistlines expand and their risk of type 2 diabetes increases, they still don't regret their past decisions, because the trade-off is optimal for them.

These individuals (if they exist) are likely to represent a tiny minority of the almost 2 billion overweight and obese adults on the planet. A more likely explanation for the obesity epidemic is expressed in Dodd (2008):

"When planning for future actions an individual may optimally wish to choose a healthy diet and lifestyle, but at every instant will be overpowered by their present bias and wish to eat unhealthily and be inactive, leaving the healthy lifestyle to start tomorrow".

Given that people *by their own reckoning* over-consume certain unhealthy foods (O'Donoghue and Rabin, 2006), and given that in most cases obesity can be influenced by lifestyle choices, including diet and exercise regimes (Dodd, 2008), it seems crucial that the obesity epidemic is tackled through policies that shift individuals towards making optimal consumption choices.

The different models offer radically different policy implications, from rational choice models that say individuals should be left alone to optimally choose obesity, to the behavioural models that encourage taxation (or discourage it), or banning advertising, or banning unhealthy goods. How can we identify which model(s) are most representative of actual human behaviour, so as to determine which policies have the best chance of curbing the epidemic?

The trouble is, all of the above models intuitively seem to contain at least a portion of the truth. It seems we do overweight present enjoyment at the expense of future health costs. There are times when the desire to eat unhealthy food does seem compulsively strong, and we do seem to focus on certain aspects of a product when making choices. Temptation does seem to play a role, as does habit, in the day-to-day choices we make.

In addition, other biases that have not yet been modelled (or where models have not gained wide-spread traction[4]) may prove to be key determinants of our consumption choices. O'Donoghue and Rabin (2015) make some suggestions as to additional causes of sub-optimal consumption, most interestingly within the observation that "economists are sometimes prone to misattribute behaviours to present bias that more likely are due to other shortcomings of the classical economics model" (p. 277). These causes include 'certainty bias' (immediate taste benefits are certain and are therefore weighted higher that uncertain future health impacts), 'anticipatory utility' (an individual experiences immediate utility from anticipating future consumption) and 'intertemporal "news" utility' (an individual experiences positive or negative utility whenever their behaviour deviates from how they thought they were going to behave).

[4] See Frederick et al. (2002) for an overview of some lesser-known models that incorporate these biases.

Naïve and Sophisticated Behaviour

O'Donoghue and Rabin (1999) separate biased consumers into two groups: naïfs and sophisticates. Naïve consumers are unaware of the biases in their decision making, and therefore falsely believe they are making (and will continue to make) optimal decisions. Sophisticates on the other hand are aware of the shortcomings in their decision-making processes, and act to counteract these failings (whether they do this successfully is another story). In particular, if a sophisticated consumer knows they will make present-biased decisions in the future, they will try and alter their behaviour, for example by committing themselves to a certain course of action in advance, to remove the option of making a poor decision in the future. Studies have shown this can manifest itself through setting deadlines (Ariely and Wertenbroch, 2002), signing up to a rolling gym membership (DellaVigna and Malmendier, 2006) or through following rules of thumb (Camerer et al., 1997)[5]. In addition, a sophisticated consumer may avoid consuming an unhealthy good at all as a means of self-control, even though it would be optimal to consume the good in moderation (O'Donoghue and Rabin, 1999).

In spite of the evidence for naivety and sophistication, the behavioural choice models listed above don't allow for naïfs to behave distinctly from sophisticates. O'Donoghue and Rabin (2015) observe (p. 275):

"While it is inconsistent with exponential discounting or sophisticated present bias for a person to predict hundreds of times that she'll start a diet, quit smoking, or write a referee report tomorrow when she won't, these seem to be types of behaviours that we observe. More and more research is suggesting that models that incorporate naivete (at least to some degree) seem to better explain behaviour."

The differentiation of individuals between naïfs and sophisticates is important from a policy perspective, as the two groups often exhibit different behaviours and can be receptive to different policy interventions. Mandel et al. (2017) propose that naïfs benefit more from commitment devices, as they tend to set higher goals, while sophisticates benefit most from 'elaboration outcomes' (where they visualise their future selves to help them improve self-control), while these are of no benefit to naïfs, who don't believe they have self-control problems.

This dissertation formulates a model where sophisticated and naïve consumers diverge in their behaviour. The model centres on internal deal-making, where consumers agree to an internal deal of under-consuming in the future in exchange for excess consumption in the present. The deal has the dual aims of satisfying immediate cravings and achieving good health outcomes *on average*, but the inability to stick to this commitment lowers life-time utility. Sophisticated consumers recognise that they won't be able to follow through on their soft commitment and adjust their behaviour accordingly, while naïve consumers believe they will.

The Scope of this Dissertation

The dissertation begins with a critical review of the key concepts, assumptions, policy outcomes and shortcomings of existing behavioural models, before examining the present-bias

[5] These studies show that neither sophistication nor naivity is consistently better: sophisticates outperformed naifs in the procrastination study (people who set regularly-spaced deadlines got better grades than people who didn't), but underperformed in the gym membership study (people who committed to a monthly membership spent more money per session than if they'd attended casually) as well as in the taxi study (taxi drivers who worked until they earned a set amount of money every day earned less overall AND had less time off than taxi drivers who worked a fixed number of hours a day).

model in depth, in order to ground the reader in the context of behavioural choice, provide key background information, and allow for a comparison of assumptions and policy outcomes with the proposed model.

A model of internal deal-making is then formulated, and the welfare impacts, applicability, and policy implications of the model are analysed. Behaviours of naïve and sophisticated consumers are predicted.

Finally, the model is compared to existing behavioural models, and the real-world applicability is discussed.

Chapter 2: Literature Review

This Literature Review has three parts. The first sections below provide an overview of behavioural decision-making models, including key assumptions and policy implications.

The section entitled 'Welfare and Decision Making' examines the split between welfare and choice that arises when rational choice assumptions are abandoned, and the different philosophical and practical efforts that have been made to solve this problem. An example model of potato chip consumption is presented, and provides concrete modelling of the concepts and challenges in evaluating choice and welfare.

Finally, the section 'Present Bias in Decision Making' specifically examines present bias as a cause of sub-optimal consumption decisions. This bias is highlighted as it has gained the most traction in the literature, and a lot of work has been done to 'price out the bias' via optimal taxation policies. The 'standard' present bias model forms the basis for the model of internal deal-making proposed in this dissertation, and a fuller understanding of the present bias model is useful in that respect.

2.1 Models of Behavioural Decision Making

Models of behavioural decision making (as opposed to the standard 'rational' decision making), have been formed in response to the evidence that a person's choices may not maximise their own welfare. Individuals exhibit a range of biases that distort their consumption decisions away from optimality, and a number of models have been produced to attempt to reflect these biases, and to reflect the extent of resultant distortion away from optimality.

The bias most commonly accounted for in these models is *present bias*[6] (see, for example O'Donoghue and Rabin, 2003 & 2006, Farhi and Gabaix, 2020, Allcott et al., 2019), which reflects our desire for immediate gratification, such that we attach disproportionate weight to immediate benefits/costs relative to costs and benefits that only occur in the future. With present bias, an individual will overconsume unhealthy food because they will overweight the (immediate) taste benefits, and underweight the (future) health costs.

Gruber and Köszegi (2001) go a step further and combine present bias with addiction (the tendency for past consumption to influence present consumption) in their model.

People often have misunderstandings about the health impacts of unhealthy foods, either through incorrect beliefs (about sugar or fat content for example) (Bollinger et al., 2011) or salience. Lockwood & Taubinsky (2017) give the example that: 'A sugary ice cream that is advertised as "fat free" may appear healthy to consumers who did not examine the less conspicuously displayed information on sugar content.' People may also be subject to misinformation, for example through conflicting messages about the dangers of certain food types, diets, etc (Gostin & Gostin, 2009).

Behavioural models have included these misunderstandings, mainly by relaxing the assumption that sub-optimal consumption is due to present bias alone, without fundamentally changing the core model (see, for example Allcott et al., 2019).

Other models base sub-optimal consumption decisions on different biases.

[6] A full explanation of present bias, how it influences decisions and how it is commonly modelled is presented on page 21.

Bernheim and Rangel (2004), postulate the existence of 'hot' and 'cold' states, where a consumer has a compulsive urge to consume a substance in the hot state, but makes a rational choice whether or not to consume in the cold state. The consumer enters the hot state when exposed to environmental cues, but has some control over the likelihood of encountering these cues by choosing which activities to participate in (we're less likely to be triggered into eating unhealthy food while going for a run than attending a child's birthday party where people are handing out slices of cake). The model also includes previous levels of consumption as a determining factor in current consumption.

Loewenstein et al. (2003) examine projection bias as a source of sub-optimal decision making, where people consistently under-appreciate the degree to which their tastes will change over time. The model then adds additional assumptions around habit formation, which results in the conclusion that consumption of the good increases over time.

Köszegi and Szeidl (2013) construct a model of 'focusing', where individuals overweight aspects of a product where the benefits/costs are concentrated, or where the product has one clear advantage over others (as opposed to when the advantages and disadvantages are more subtle and need to be evaluated across many aspects). The conclusion of this model is that present bias exists only in certain circumstances. In the case of unhealthy foods, the taste enjoyment occurs immediately, whereas the health impacts are diffuse and occur over many periods. This means that a consumer is likely to be highly sensitive to price changes of an unhealthy item, as payment is also something that occurs immediately, and is therefore an aspect of the choice that the individual is likely to focus on.

Gul and Pesendorfer (2001) include temptation in their model, which postulates a utility loss associated with not choosing the most tempting option. In this model, the consumer chooses optimally, as it is better to suffer adverse health consequences than to have to deal with the loss of utility from not eating the most tempting option. This model allows for the possibility of commitment by the individual. For example, if the individual knows they will be tempted by a hamburger if it's offered on a menu, they can choose to go to a restaurant that doesn't offer hamburgers, thereby avoiding temptation, which can otherwise only be resisted through costly self-control.

2.2 Policy Implications of Behavioural Models

In earlier present bias literature (O'Donoghue and Rabin 2003 & 2006, for example), a 'sin tax' is proposed to counter the bias and return the user to optimality. Although sin taxes may provide benefits in terms of allowing individuals to consume more optimally (by consuming less unhealthy food and thereby improving their health), this comes at a financial cost. Given that poor people are more likely to overconsume unhealthy food[7], sin taxes are financially regressive and disproportionately borne by the poor, even though they may be progressive from a health perspective (Harkanen, et al., 2014). More recent studies (see, for example, Kotatorpi, K., 2008, Allcott et al., 2019) propose a trade-off between *corrective benefits* and *regressivity costs* (Lockwood & Taubinsky, 2017), such that a sin tax should be used when the consumption of the good is highly sensitive to price, but other non-tax methods should be used to reduce consumption when consumers are not highly responsive to price (Farhi and Gabaix, 2020). These non-tax methods include 'nudges,' which change the framing of a decision without changing the financial incentives or restricting choices (Bernheim and Taubinsky, 2018).

Rather than the lump-sum redistribution of taxation revenue used in O'Donoghue and Rabin (2006), Allcott et at. (2019) use a non-linear (progressive) income tax to return sin tax revenues to consumers.

Farhi and Gabaix (2020), extend this work on different taxation mechanisms and influences when acknowledging behavioural biases by including tax salience as an additional factor in modelling (i.e. accounting for individuals not being fully attentive to the tax component of their purchases and income).

The Koszegi and Szeidl (2013) model of focusing also encourages taxation as a means of shifting the individual towards optimal behaviour. This is because price is immediate and salient, and is something an individual is likely to focus on and be sensitive to when weighing up benefits and costs in a consumption decision.

In the Bernheim and Rangel (2004) model of hot and cold states, consumption in the hot state is considered to be largely unresponsive to price changes. Consumption in the cold state is responsive to price, but cold-state consumption is assumed to optimal, so deviations in the cold state due to price changes would lower utility. In the hot state, a higher price would not change consumption, but would lower utility through forcing the individual to pay a higher price. Any optimal tax is therefore negative (i.e. the good should be subsidised). In this model, environmental triggers cause the individual to enter the hot state, so an effective policy would be to remove these triggers, through banning advertising, for example. In the present bias model presented above, banning advertising would have no impact on consumption, as advertising is not assumed to impact consumption decisions. Note that empirical studies have shown that sugary drink consumption (for example) does respond well to price increases (Allcott et al., 2019), so this model of hot and cold states may not have strong real-world applications, despite its intuitive appeal, and the presence of visceral influences that is missing from other models.

A key difference in the Gul and Pesendorfer (2001) model of temptation is that the individual *is* consuming at an optimal level, because consuming otherwise would result in a utility loss from not having chosen the most tempting option. The trade-off between future health impacts and immediate enjoyment is weighted not only by the taste of the chosen option, but also from

[7] See, for example pp 1562 of Allcott et al. (2019) for a breakdown of household sugary drink consumption by income level.

the immediate temptation disutility when the most tempting option isn't chosen. Finding the optimum level of consumption in this circumstance means less weight is given to future health, which results in poorer health outcomes. Although a tax would likely reduce consumption, the most obvious policy outcome is the removal of temptation (either directly, or by allowing the consumer to remove temptation themselves through commitment options). Once the unhealthy food is removed as an option, the person's overall utility increases, as they can enjoy better health without suffering temptation disutility as a consequence.

It is interesting to consider whether individuals would vote for the implementation of the above policies (assuming the policies had been correctly identified as leading to the intended behavioural change). A sophisticated individual (who knows they suffer from a behavioural bias) is likely to vote for such policies, as they provide a way of changing behaviour that the individual would be unable to achieve on their own. Naïve individuals believe they already consume optimally, and would normally argue against such policies (or at most be indifferent), as they (incorrectly) believe the policies would harm their welfare. However, O'Donoghue and Rabin (2003) argue that in some instances naïfs would also vote in favour of sin taxes, as these taxes, when combined with lump-sum transfers can be welfare improving (at least on average) for rational consumers (which naïve consumers believe themselves to be).

2.3 Welfare and Decision Making

It is generally accepted (both by economic professionals and laypeople) that people try to make decisions that lead to the best outcomes for them. We are all unique, with different tastes and priorities, therefore we each make different decisions, but we are all trying to make the decisions that are best for us, and that lead to improved welfare. All else being equal, a person who loves potato chips will consume more of them, and a person who places a high value on health will consume less unhealthy food.

This 'truth' can be approximated mathematically:

A basic model of potato chip consumption

For an individual who enjoys eating potato chips, the daily taste utility from potato chip consumption can be modelled as:

$$u^{taste} = \frac{a}{1-r}x^{1-r}$$

Where $\frac{a}{1-r}$ is the taste utility derived from the first potato chip of the day, and the decreasing marginal enjoyment of subsequent potato chips follows the path x^{1-r}, where x is the number of potato chips eaten, and r takes a value between 0 and 1. The larger the value of r, the faster the marginal enjoyment declines.

Unfortunately, eating potato chips comes with consequences. Every potato chip eaten increases the number on the scales the next morning, and increases the likelihood of developing type 2 diabetes and other adverse health effects.

For this model, let's assume the health impacts of potato chips are proportional to consumption[8]: for each unit of potato chips consumed, health gets b units worse.

The health impact of potato chips is therefore:

$$u^{health} = -bx$$

In addition, individuals have limited incomes, and have to decide how to split their income between expenditure on potato chips, and expenditure on everything else in their lives.

If I in an individual's income, and z is everything they could spend their money on that isn't potato chips, then:

$$I = p_x x + p_y z$$

In this formula, the units of each good (x and z) are normalised so they have identical marginal costs, which in a competitive market also means identical prices. If the price of good x is itself normalised to 1, this removes the need to explicitly add prices to the above formula:

$$I = x + z$$

[8] This is almost an arbitrary assumption, and empirical evidence may prove that it is incorrect. If health impact had another functional form (if health impacts increased exponentially with consumption, for example), the exact optima would of course be different, but the underlying patterns and conclusions derived here would remain unchanged.

The utility associated with the 'everything else' good is assumed to be linear and unitary (i.e. the consumption of one unit of good z increases an individual's utility by 1). These assumptions simplify the equations used in the model, without reducing their applicability.

The net utility associated with everything you consume today is therefore:

$$u^{net} = u^{taste} + u^{health} + u^{everything\ else}$$

$$u^{net} = \frac{a}{1-r}x^{1-r} - bx + z$$

Since z can be written as:

$$z = I - x$$

This gives:

$$u^{net} = \frac{a}{1-r}x^{1-r} - bx + I - x$$

This specific form of potato chip utility is taken from O'Donoghue and Rabin, 2006.

Note that not all components of this utility function occur immediately. Taste utility and the utility gained from consuming everything else is likely to be immediate and experienced at the time of consumption, whereas health disutility is likely to be delayed and only occur some time in the future. This distinction will become important later.

Given that eating potato chips increases utility by providing deliciousness, and decreases it by coating our insides with excess fat and salt, what is the optimal number of potato chips to consume in a day?

The value of potato chips (x), that maximises the value of u^{net} is:

$$\frac{du}{dx} = ax^{-r} - b - 1 = 0$$

$$x = \left(\frac{a}{b+1}\right)^{\frac{1}{r}}$$

The Divergence of Choice and Welfare

According to rational choice theory, this is exactly what happens. We spend our days wandering around, (implicitly) solving differential equations and consuming whatever the solution to the differential equation tells us to consume. For rational choice theory, choice IS welfare. If we make a choice, it is by definition a manifestation of us optimising our welfare.

Unfortunately, the evidence suggests that in reality we aren't very good at doing this (see, for example, Frederick et al., 2002), and we suffer from a large range of behavioural biases that get in the way, and cause us to make decisions that don't optimise our utility (i.e. we spend our days wandering around, solving the *wrong* differential equations and thereby consuming sub-optimally).

This has some serious implications. There are obvious policy consequences; instead of leaving people alone to make 'optimal' choices, the situation now calls for interventions to help improve welfare beyond what the individual could achieve on their own. When determining the consequences of a given policy, or when seeking to implement a policy to achieve a certain aim, standard assumptions around rationality can result in misleading conclusions (Bernheim and Taubinsky, 2018).

The solution is unfortunately more complicated than simply abandoning rational choice assumptions and replacing them with assumptions that better predict observed behaviour. If we agree that choice isn't welfare, then what is?

Before the emergence of the field of behavioural economics, the question of 'what is welfare?' wasn't very important. People made choices that optimised their welfare, and so long as they were left to it, the definition of what exactly they were optimising didn't really matter.

But now the situation is a lot more complicated. Welfare is no longer related to something that we can see, touch or measure. Individuals don't manifest it in the choices they make, it is hidden, and yet it is what we must strive to optimise if we want to make people better off.

As Bernheim and Rangel (2007a) state: "The fundamental problem of behavioural welfare economics is to identify appropriate criteria for evaluating alternatives when, due to choice reversals and other behavioural anomalies, the individual's choices fail to provide clear guidance."

Bernheim and Taubinksy (2018) summarise the challenge as: 'the need to maintain the tight link between empirically measurable statistics and welfare estimates, while moving beyond the revealed preferences assumption'.

In spite of the challenges, it is necessary to meet them. Standard welfare analysis is based on choice, *not* on utility or preferences (Bernheim and Rangel, 2007a). Since in reality choice often doesn't correspond with welfare, policies that seek to optimise choice have at best only an arbitrary impact on welfare, and at worst can be detrimental.

Behavioural economists have postulated that there exists a 'true' utility function (first defined as 'experienced' utility in Kahneman, 1994), that represents the utility that actually optimises our welfare.

Unfortunately, because of our many behavioural biases, when we make decisions, we seek to optimise a different utility function, called our 'decision' utility:

For example, if we suffer from present bias or focusing or certainty bias, we might overly discount the health impacts of potato chip consumption because they only occur in the future or because they are diffuse or uncertain.

The impossibility of measuring 'true' welfare

It is easy to define and obtain empirical evidence for decision utility: it is simply whatever the individual chooses. The definition of 'true' utility, on the other hand, is not a straightforward exercise.

In the case of present bias, the individual, when thinking ahead about the future, sets out a consumption plan that corresponds with (presumably) bias-free decision making. However, when the future arrives, the individual consumes a larger amount than they had planned, because present bias causes them to overweight present utility and underweight future health costs (which is not an issue when comparing utilities that only occur in the future[9]). Tomorrow, when thinking back on the over-consumption that occurred today, the individual will regret their consumption decisions, and wish that they had consumed in accordance with their long-term plans.

[9] See discussion on time-inconsistency and quasi-hyperbolic discounting on page 21.

This is the justification why, in the case of present bias, the long-term perspective is considered to represent 'true' utility, while the immediate perspective is considered to represent the biased 'decision' utility. Proponents of present bias claim that this is an uncontroversial position to take. O'Donoghue and Rabin (2006), for example, state that their proposed true utility function 'is appropriate under essentially any perspective', and that 'for any tax policy that takes effect in the future… the agent agrees that [it] is the appropriate welfare function.' i.e., the person's present bias reflects "a short-term desire or propensity that the person disapproves of at every other moment in her life." (O'Donoghue and Rabin, 2003, p. 187).

Farhi and Gabaix (2020) claim that 'choices in environments where behavioural biases are attenuated can be thought of as rational,' and specify these environments as including 'if agents have a lot of time to decide, taxes and long-run effects are salient, and information about costs and benefits is readily available.'

However, concluding that long-term preferences are 'correct' and short-term preferences represent errors represents a moral judgement call on the part of the policy maker. Bernheim (2016) points out that many cultures emphasize the importance of living in the moment. In addition, Lockwood and Taubinsky (2017) mention that our future-oriented selves suffer from an over-abstraction bias, which would indicate that the long-term view is not bias-free. Köszegi and Szeidl (2013) identify instances when individuals can suffer from 'future bias', for example when planning for big events, and, if asked to commit to future actions, we will 'over-commit' to beneficial activities, for example to eating well, exercising, etc. The welfare dominance of long-term over short-term preferences is therefore not obvious.

Allcott et al. (2019) suggest policy could incorporate a normative weighting between the 'future-oriented self' and the 'in-the-moment self.'

Bernheim, & Rangel, (2007a) provide a framework where no set of preferences is assumed to be the 'true' set, and concludes that only changes that result in welfare improvements across *all* sets of preferences can be assumed to be welfare improving. However, practically this does not offer a way forward in many instances. For example, if the obesity epidemic is caused by present biased decision making, but we are unable to improve long-term utility if it leads to a loss of short-term utility (i.e. if we can't improve health without restricting a person's day-to-day consumption of unhealthy food), then we don't have a lot of options left to play with.

Models Incorporating 'True' and 'Decision' Utility

Despite the difficulties in defining true utility, it is still a necessary feature of models, as it represents the optimal welfare path.

In our potato chip example, true utility is assumed to be:

$$u^{true} = \frac{a}{1-r}x^{1-r} - bx + I - x$$

While decision utility is assumed to be some distorted version of this true utility function.

For example, if we suffer from present bias or focusing or certainty bias, we might overly discount the health impacts of potato chip consumption because they only occur in the future or because they are diffuse or uncertain.

If β (with a value $\beta < 1$) is a discount factor arising from our bias and is applied to any utility component that only occurs in the future, then our decision utility is:

$$u^{decision} = \frac{a}{1-r}x^{1-r} - \beta bx + I - x$$

(again, this representation is taken from O'Donoghue and Rabin, 2006).

This gives an optimal (true utility based) potato chip consumption of:

$$x^{true} = \left(\frac{a}{b+1}\right)^{\frac{1}{r}}$$

And an actual (decision utility based) consumption of:

$$x^{decision} = \left(\frac{a}{\beta b+1}\right)^{\frac{1}{r}}$$

In order to examine the impacts of this graphically, let's give our parameters some numerical values. Let's assume:

$$a = 10 \qquad b = 2$$
$$r = 0.5 \qquad \beta = 0.9$$
$$I = 50$$

These values are arbitrary, and don't make any claim on representing reality. To see empirical studies that attempt to determine actual parameters (with a different focus than the exercise presented here), see Allcott et al. (2019), Harkanen et al. (2014), Dharmasena and Capps (2012) or Briggs et al. (2013).

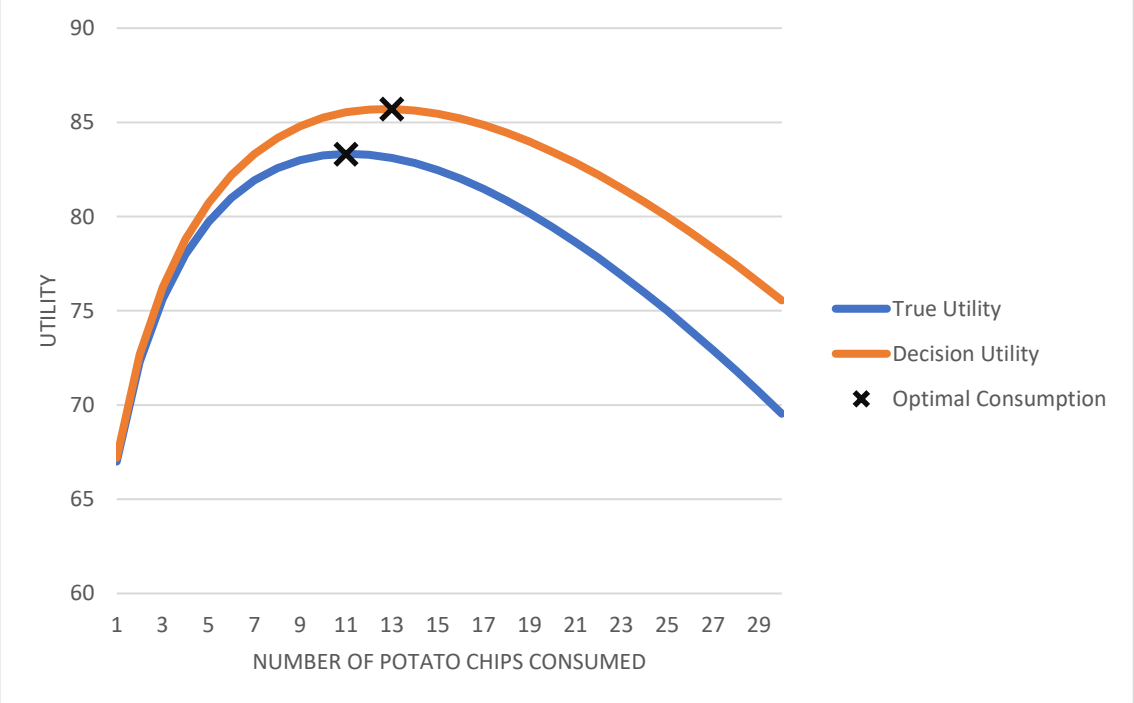

Figure 1: The Divergence of Decision Utility and True Utility in Potato Chip Consumption

This graph shows that the optimal quantity of potato chips (the quantity at which the net utility is highest) is greater for decision utility than for true utility. In reality, the individual should consume 11 potato chips a day, however, the individual's biased decision making causes them to consume 13 potato chips, which gives them a lower true utility than if they'd been able to act without bias distorting their decisions.

This may seem unimportant, but when viewed through the lens of the obesity epidemic, the divergence of decision utility from true utility and the resultant over-consumption has serious consequences.

Measuring Welfare Loss

Just how bad is it if an individual consumes 13 potato chips a day rather than 11? If we return to the real world and observe the impacts of the obesity crisis, it seems the answer is "pretty bad."

A measure that has become common in the literature (see Chetty et al., 2009, Allcott et al., 2019, Farhi & Gabaix, 2020) is the 'price metric of bias', which is a measure of the price change required to "price out the bias" (Bernheim and Taubinsky, 2018). It is the price change that would cause non-biased optimal consumption to equal biased consumption (at the original price). In our example:

Explicitly including potato-chip price p, and using a price-metric of bias γ, un-biased utility is modelled as:

$$u^{true} = \frac{a}{1-r}x^{1-r} - bx + I - px$$

Which has an optimum at:

$$x^{true} = \left(\frac{a}{b+p}\right)^{\frac{1}{r}}$$

Setting this equal to the biased optimum and solving for γ :

$$x^{decision} = \left(\frac{a}{\beta b + p}\right)^{\frac{1}{r}} = \left(\frac{a}{b + (p+\gamma)}\right)^{\frac{1}{r}}$$

Yields:

$$\gamma = \beta b - b$$

When $b = 2$ and $\beta = 0.9$ (as in our example), $\gamma = -20c$

This means that a 20% per-chip price reduction (remembering that the price is normalised to \$1) causes optimal consumption to shift from 11 chips to 13 chips, which is the same consumption shift caused by our bias.

2.4 Present Bias in Decision Making

Present bias affects an individual's potato chip consumption because the associated health impacts only occur in the future, while the consumption decision and positive taste utility are immediate.

To an extent, discounting future utility is logical and rational, and is a form of optimisation. If there is a non-zero chance that future utility impacts won't occur, they should be discounted proportionately. For example, if the individual dies in an accident, or there is a global famine, then the negative future consequences of excess potato chip consumption will not arrive. The further into the future the negative health impacts are expected to occur, the more they can be discounted, as there is a greater probability they will never arrive. Equally, the individual may simply 'get lucky' and not have to face any consequences, as in reality there is no guaranteed cause-and-effect relationship between consumption and health costs (i.e. and relationship between unhealthy eating and poor health is true only on average, and any given individual will vary in their actual health consequences).

Other interpretations of future discounting include thoughts around earning interest ($100 now is better than $100 later, as interest can start to accumulate as soon as the $100 is received). Discounting in this context is proportional to the rate at which interest can be earned.

No matter the reasoning, the discounting of future utility is seen as being part of human nature[10].

In our potato chip example this 'exponential discounting' is modelled under the form:

$$U^{true} = \sum_{t=1}^{T} \left(\frac{a}{1-r} x_t^{1-r} - b x_{t-1} + I - x_t \right) \delta^{t-1}$$

Where U^{true} is lifetime utility over T periods, x_t is potato chip consumption in period t, x_{t-1} is potato chip consumption in period $t-1$ (we're assuming that health impacts from consumption in period t are experienced in period $t+1$), and δ is the exponential discount factor, which has a value $\delta \leq 1$.

This means that, when we're standing at time $t = 0$ and looking forward, our projected utility for a given event will decrease the further out into the future it occurs. For example, consuming ten potato chips today will give us a taste utility of:

$$u^{taste} = \frac{a}{1-r} 10^{1-r}$$

Recall from the numerical example above:

$$a = 10 \qquad\qquad r = 0.5$$

This gives a taste utility (ignoring any health consequences or other consumption goods) of 63 utils. However, if we assume an exponential discount factor of $\delta = 0.95$, then projecting forward, that same ten potato chips corresponds to a smaller and smaller amount of taste utility, depending on when they are consumed:

[10] For an overview of the different interpretations of future discounting see Dimitri and van Eijck (2012).

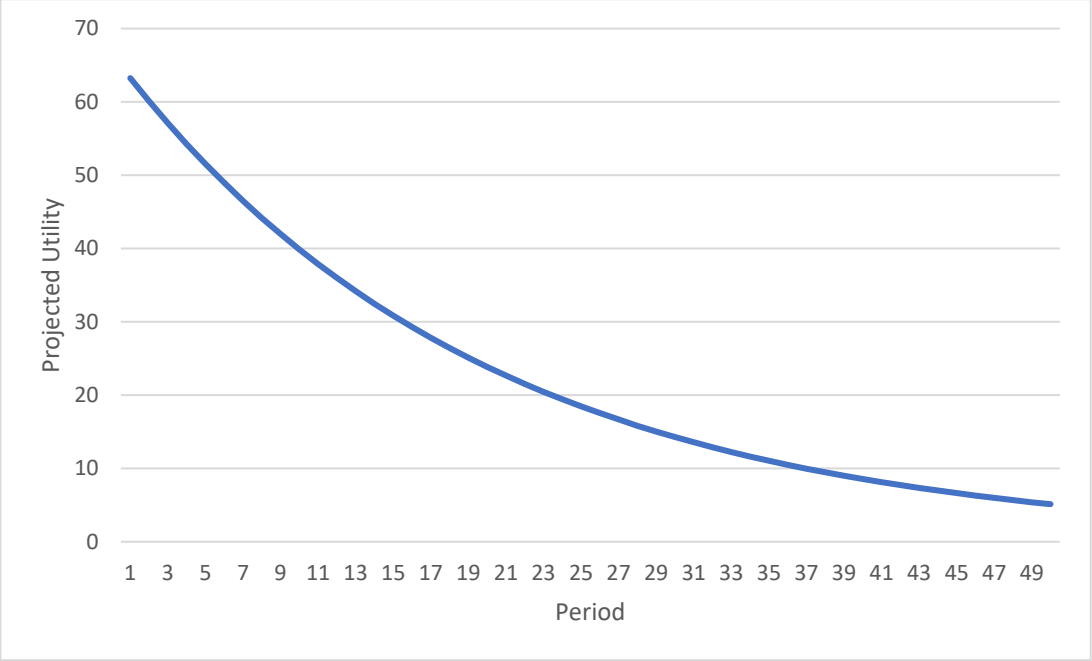

Figure 2: Exponential Discounting of Future Utility

This form of discounting is *time consistent*, in that relative preferences do not change over time (i.e. the individual feels the same about the relative benefits of an event occurring today or occurring next week as they do about the relative benefits of an event occurring next week versus the week after).

However, experimental evidence (see, for example Benzion et al., 1989) suggests that individuals, when discounting future utility, don't actually follow the exponential-type discounting modelled in Figure 2.

An intuitive example of how people actually discount the future is given in O'Donoghue and Rabin, 1999 (p. 103):

> "When presented a choice between doing seven hours of an unpleasant activity on April 1 versus eight hours on April 15, if asked on February 1 virtually everyone would prefer the seven hours on April 1. But come April 1, given the same choice, most of us are apt to put off the work until April 15."

This is an example of *time inconsistency*, where our relative preferences differ over time.

Evidence suggests this time inconsistency is a result of our present bias, where instead of a progressive discounting of future periods, we overly discount all periods that aren't today (i.e. we exhibit a strong bias towards utility that occurs *right now*). Laibson (1997), proposed a model of quasi-hyperbolic discounting to account for this effect. Quasi-hyperbolic discounting in our potato chip example has the form:

$$U^{decision} = \frac{a}{1-r}x_1^{1-r} + I - x_1 + \beta \sum_{t=2}^{T} \left(\frac{a}{1-r}x_t^{1-r} - bx_{t-1} + I - x_t \right) \delta^{t-1}$$

The difference between exponential and quasi-hyperbolic discounting is the presence of the β term, which adds an extra discount factor to all periods that aren't today, and leads to time inconsistency.

In comparison to the exponential discounting of taste utility above, with quasi-hyperbolic discounting, if $\beta = 0.9$, then standing at time $t = 0$ and looking forward, our taste utility over time for those ten potato chips now has the form:

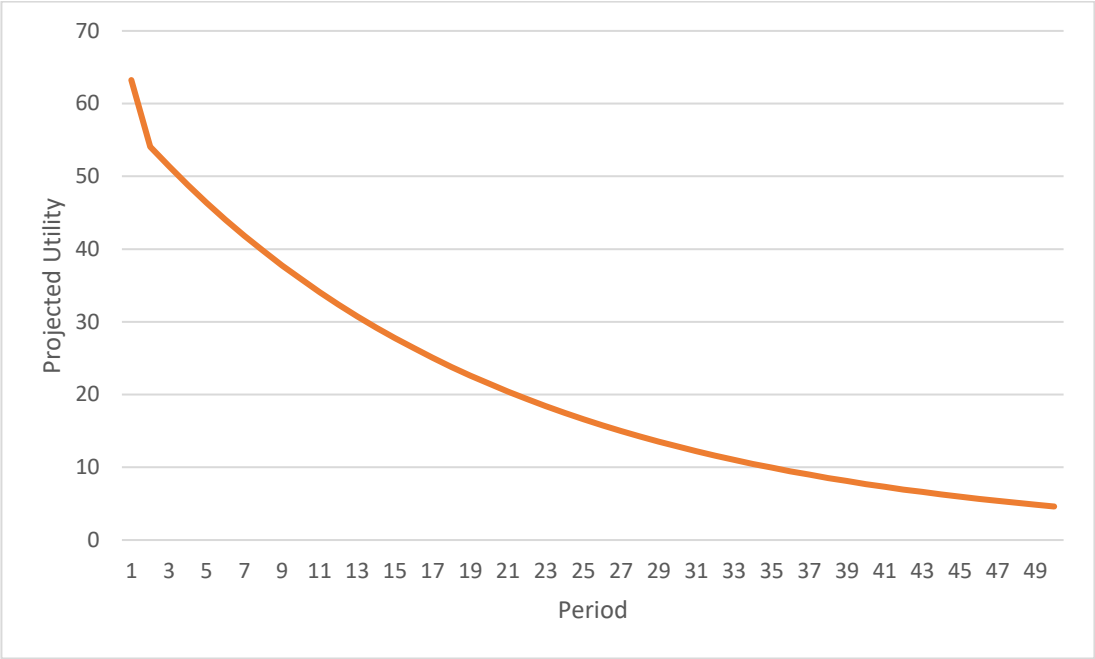

Figure 3: Quasi-Hyperbolic Discounting of Future Utility

Present bias still exhibits normal exponential discounting when comparing one future date with another, but shows a disproportionately larger discount between today and any future date. Because of the extra discounting applied to all periods that aren't today, when weighing up a consumption decision between today and tomorrow, an individual could come to a different conclusion than when weighing up the same decision between tomorrow and the day after. This means that individuals can struggle to follow through with plans they've made. For example, people often make plans to start eating well or exercising in the future, but as we've all experienced, these plans can be hard to stick with once the future arrives and the healthy eating and gym-going actually has to happen.

Present bias/time inconsistency is the bias most commonly accounted for in the literature when modelling departures from optimal consumption decisions.

The 'Standard' Model of Present Bias

O'Donoghue and Rabin (2006) proposed a model of present-biased consumption decisions, based on quasi-hyperbolic discounting. This model has gone on to form the basis (or at least the inspiration) of many present bias models of intertemporal consumption decisions[11]. In addition, other biases that aren't present bias, for example limited attention/salience, misinformation and certain forms of focusing, can also be modelled using this present bias

[11]The literature often presents a model of present biased consumption in more generic terms (see, for example Farhi & Gabaix, 2020, Allcott et al., 2019, and even O'Donoghue and Rabin, 2006), for example by allowing for unspecified functional forms of the taste utility and health (dis)utility, allowing for more consumption goods and a more generic form of bias. However, when called upon to extract a specific form of that model for quantitative analysis, it inevitably bears a strong resemblance to the model proposed in O'Donoghue and Rabin (2006).

model, as the impact on behaviour is the same, in that they all result in discounting of future health costs.

The O'Donoghue and Rabin (2006) model uses quasi-hyperbolic discounting to account for present bias, and assumes health impacts occur in the period immediately following consumption. An investigation of the actual form of the health impact is likely necessary in empirical studies, but the simplified assumptions used here don't impact the overall trends and implications of the model. The (biased) decision utility and (unbiased) true utility over time are:

$$U^{decision} = \frac{a}{1-r}x_1^{1-r} + I - x_1 + \beta \sum_{t=2}^{T} \left(\frac{a}{1-r}x_t^{1-r} - bx_{t-1} + I - x_t\right)\delta^{t-1}$$

$$U^{true} = \sum_{t=1}^{T} \left(\frac{a}{1-r}x_t^{1-r} - bx_{t-1} + I - x_t\right)\delta^{t-1}$$

For simplicity, it is assumed $\delta = 1$.

In this model, all benefits and costs from consumption in a given period are additively separable from all other periods. This means that consumption decisions in one period are independent of consumption in other periods, and in every period the individual faces an identical consumption decision.

In each period, the individual's present bias causes them to consume by optimising their decision utility:

$$u_t^{decision} = \frac{a}{1-r}x_t^{1-r} - \beta b x_t + I - x_t$$

Which results in a consumption decision of:

$$x_t^{decision} = \left(\frac{a}{\beta b + 1}\right)^{\frac{1}{r}}$$

But the utility they *should* be optimising (no present biased discounting) is:

$$u_t^{true} = \frac{a}{1-r}x_t^{1-r} - b x_t + I - x_t$$

Which corresponds to a bias-free consumption level of:

$$x_t^{true} = \left(\frac{a}{b + 1}\right)^{\frac{1}{r}}$$

Present bias increases consumption of unhealthy food relative to the optimal, bias-free level.

As we know from earlier, this corresponds to a price metric of bias of:

$$\gamma = \beta b - b$$

O'Donoghue and Rabin (2006) propose an 'optimal sin tax' to counteract this over-consumption. Subsequent literature has continued work in this direction, making taxation mechanisms and calculations more comprehensive and nuanced. The underlying present bias model, however, has not changed greatly since 2006 (except perhaps to become more generic).

Present bias, it seems, has been accepted as the underlying bias that causes us to deviate from optimal consumption.

However, evidence indicates that consumption 'mistakes' are actually triggered by intermittent environmental cues, while present bias is modelled as *always* present (Bernheim and Rangel, 2007b). Köszegi and Szeidl (2013) propose that the extent of present bias differs depending on the situation, and is lower when future costs are less dispersed, or when a person commits to fewer decisions with accumulating effects on their planned consumption. Given that O'Donoghue and Rabin (2015) suggest that many behaviours assumed to be a consequence of present bias are actually due to other factors, it is not obvious that present bias models are appropriate to derive effective welfare-enhancing policies.

Visceral Influences and Restraint Bias

Present bias is modelled as a 'constant pull'. Its effect is always working on us, continually distorting our consumption decisions away from optimality. However, in reality we are often perfectly capable of making appropriate decisions. Impulsive decisions that deviate from optimality arise when we are tired, hungry, aroused, bored…i.e. they are the result of visceral influences that overwhelm our ability to consume at what we would otherwise consider to be an optimal level.

Loewenstein (1996) proposes that visceral factors increase the perceived immediate value of a good, while having little impact on its perceived long-term value. The Bernheim and Rangel (2004) model assumes that visceral influences trigger a person to enter a 'hot' state, where they are insensitive to price changes and have a compulsive urge to consume the unhealthy good.

Hoch and Loewenstein (1991) claim that time inconsistency is in fact a battle between desire and willpower.

In spite of these visceral influences making us want to overconsume *now*, we seem to exhibit a remarkable degree of optimism regarding our ability to overcome the same visceral forces in the future.

Loewenstein (1996) proposed that individuals underestimate the impact of visceral factors on their own future behaviour. Although we are currently tired and hungry and reaching for those potato chips, we believe the future will be different, that tomorrow we will not be prone to these temptations, either because we don't expect to find ourselves hungry, tired and stressed, or because we believe we will easily be able to resist the urge to over-consume unhealthy food when subject to visceral influences in the future.

This over-optimism about our ability to resist future temptation is known as 'restraint bias' (Nordgren et al., 2009).

Chapter 3: Deal-Making and Soft Commitment

This section formulates an alternative model, centred on the tendency individuals have to make deals with themselves to justify overconsumption of unhealthy foods, and the restraint bias that makes us overly-optimistic about our ability to stick to these deals.

Present bias implies that we are unable to properly weight future health impacts. The model presented here claims that we *are* able to accurately weight health impacts, and instead it is the mis-exploitation of 'wiggle room' that causes us to deviate from optimal consumption.

A Model of Internal Deal-Making and Soft Commitment

In this model, individuals fully understand and appropriately weight the health impacts of unhealthy food consumption (unlike in the case of present bias), and therefore have no problem calculating optimal consumption levels.

Using the basic model of potato chip consumption set out above, this means:

$$u_t = \frac{a}{1-r}x_t^{1-r} - bx_t + I - x_t$$

Here, u_t represents the utility *derived from* consumption in period t, as opposed to the utility that *occurs in* period t. This means that the health impacts are also assigned to period t, as this is when they are instigated. In this model it is not important when the health impacts associated with today's potato chip consumption occur, which also means the health impacts could occur over multiple periods (i.e. potato chip consumption today can impact health over many future periods).

The consumer has a rational optimal consumption of:

$$x_t = \left(\frac{a}{1+b}\right)^{\frac{1}{r}}$$

The individual is fully aware of this optimal level of consumption. However, the individual is also vulnerable to visceral influences, which makes them 'obsess' about the immediate taste utility they could obtain from eating potato chips RIGHT NOW. These visceral influences (hunger, boredom, stress, social pressure…) make the individual heavily discount other sources of utility, including any health consequences of current potato chip consumption.

The visceral side of us is not always present (there are times when we have no trouble eating in accordance with what we know we should eat). However, when our visceral side does emerge, it *really* wants to eat potato chips.

The visceral side of us wants to optimise:

$$u_{visceral} = \frac{a}{1-r}x_1^{1-r} - \theta bx_1 + \theta(I - x_1)$$

Where θ is the discount factor, and has a value between 0 (which corresponds to completely disregarding non-taste utility) and 1 (no discounting). In reality θ is likely to have a value close to zero, i.e. the visceral side of us *strongly* discounts any utility that isn't related to immediate potato chip consumption. The visceral side of us is uninterested in any health consequences,

and is equally unmoved by the knowledge that consumption of carrots (or some other non-potato chip good) can also provide us with immediate utility[12].

This visceral utility has an optimal consumption of:

$$x_{visceral} = \left(\frac{a}{\theta(1+b)}\right)^{\frac{1}{r}}$$

The individual is now in conflict between how many potato chips they know they *should* consume and how many they *want* to consume right now.

The individual now faces a choice: they can ignore their visceral cravings and consume rationally, they can ignore their rational knowledge and consume viscerally, or they can consume at some other level.

Note that in other behavioural models, the definition of 'true' utility was externally imposed by the economist, i.e. the economist made a judgement call as to whether the short-term or long-term self was correct, or whether some relative weighting should be given to each self. In this model, however, the individual themselves gets to decide which self gets precedence, and can (attempt to) find a compromise between their visceral and rational selves, however they see fit.

In the case of unhealthy foods, when we are subject to visceral forces, our rational selves are screaming out for us to think about our health, all while our visceral selves are wanting us to over-consume. In these moments, our brains are frantically searching for a way to keep us healthy all while eating more potato chips than we know we should.

In order to maintain the same level of health, this optimal consumption only needs to be achieved *on average*. For example, over two periods, the optimal level of potato chip consumption is:

$$x_1 + x_2 = 2\left(\frac{a}{1+b}\right)^{\frac{1}{r}}$$

The same health benefits can be achieved by consuming $\left(\frac{a}{1+b}\right)^{\frac{1}{r}}$ each period, but can also be achieved by 'trading-off' between periods, for example by over-consuming one period, followed by under-consuming the next. Any consumption combination is possible, provided the consumption over both periods sums to $2\left(\frac{a}{1+b}\right)^{\frac{1}{r}}$.

A little wiggle room is a dangerous thing. For a person subject to visceral influences and who would really love to consume extra potato chips today, this flexibility gives them a way forward without compromising their health, provided they can stick to the soft commitment they make to themselves and consume a correspondingly smaller amount tomorrow.

Over two periods (today plus tomorrow), the consumer therefore has the following logic:

Rational self:

$$\max_{x_1,x_2} U_{1+2}^{rational} = \frac{a}{1-r}x_1^{1-r} + \frac{a}{1-r}x_2^{1-r} - b(x_1+x_2) + (I-x_1) + (I-x_2)$$

[12] The assumption that when you crave potato chips, carrots just ain't gonna cut it, is an important distinction between this representation of visceral influences and standard models of present bias, which assume all immediate sources of utility have equal weight.

Visceral self:

$$\max_{x_1,x_2} U_{1+2}^{visceral} = \frac{a}{1-r}x_1^{1-r} + \theta\frac{a}{1-r}x_2^{1-r} - \theta b(x_1+x_2) + \theta(I-x_1) + \theta(I-x_2)$$

Wiggle room for the visceral self to deviate from rational optimality:

$$x_1 + x_2 = 2\left(\frac{a}{1+b}\right)^{\frac{1}{r}}$$

The individual can therefore choose to (partially) satiate their visceral selves today, provided they under-consume by a corresponding amount tomorrow.

Objectively, the individual shouldn't accept this 'deal', as tomorrow they are just as likely as today to be subject to visceral forces and a desire to over-consume. However, as we all know from our (repeated) claims of the form "I will start eating healthy/exercising tomorrow," we are often overly optimistic about the amazing willpower and energy levels we will exhibit in the future. When we decide to over-consume today, we tend to genuinely believe that tomorrow we will have the willpower to overcome any temptation caused by the deliciousness of potato chips and the immediate taste utility they provide.

The individual therefore accepts this deal, and optimises their two-period (today plus tomorrow) visceral utility function, subject to their 'rational' restraint:

$$\max_{x_1,x_2} U_{1+2}^{visceral} = \frac{a}{1-r}x_1^{1-r} + \theta\frac{a}{1-r}x_2^{1-r} - \theta b(x_1+x_2) + \theta(I-x_1) + \theta(I-x_2)$$

$$x_1 + x_2 = 2\left(\frac{a}{1+b}\right)^{\frac{1}{r}}$$

The individual's potato chip consumption today is[13]:

$$x_1 = \left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{2}{1+\theta^{\frac{1}{r}}}$$

And planned consumption for tomorrow is:

$$x_2 = \left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{2\theta^{\frac{1}{r}}}{1+\theta^{\frac{1}{r}}}$$

Note that the solutions for x_1 and x_2 include a term that reflects average optimal consumption (highlighted in yellow), and a term that reflects deviations from that optimum (highlighted in blue). The share of consumption between period 1 and period 2 (today and tomorrow) sits between two extremes: When $\theta = 0$, the plan is to eat two-period's worth of potato chips today and no potato chips tomorrow. When $\theta = 1$, planned consumption is constant (and optimal) every day.

Many present bias models propose that in the real world the degree of discounting due to present bias is likely to be small (i.e. β values are close to 1). However, for the case of internal deal-making, any and all θ values could be possible (and values close to 0 are perhaps more likely). In internal deal-making, people could easily believe that tomorrow they will be completely impervious to the lure of unhealthy food, such that they consume 100% of their 'allowance' today, with the plan of completely abstaining from the food item tomorrow.

[13] See Appendix for derivation.

The Utility Compromise

This 'deal' to overconsume today with the promise of under-consuming tomorrow represents a compromise between the competing rational utility and visceral utility functions. Using the rational utility function as the 'true' utility (i.e. assuming the individual is best off when they optimise their rational utility), the potato chip consumption as a result of the internal-deal yields a lower utility when compared with the fully rational consumption amount:

$$U_{1+2}^{rational}\ at\ x_1 = x_2 = \left(\frac{a}{1+b}\right)^{\frac{1}{r}} > U_{1+2}^{rational}\ at\ x_1 = \left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{2}{1+\theta^{\frac{1}{r}}}, x_2 = \left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{2\theta^{\frac{1}{r}}}{1+\theta^{\frac{1}{r}}}$$

Similarly, if the visceral utility function is taken to be the 'true' utility (i.e. assuming the individual is best off when they optimise their visceral utility), the potato chip consumption as a result of the internal-deal yields a lower utility than fully visceral consumption.

$$U_{1+2}^{visceral} at\ x_1 = \left(\frac{a}{\theta(1+b)}\right)^{\frac{1}{r}}, x_2 = \left(\frac{a}{1+b}\right)^{\frac{1}{r}} > U_{1+2}^{visceral} at\ x_1 = \left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{2}{1+\theta^{\frac{1}{r}}},\ x_2 = \left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{2\theta^{\frac{1}{r}}}{1+\theta^{\frac{1}{r}}}$$

The internal-deal is the individual's attempt to trade-off between their rational and visceral utility functions, to achieve a consumption level that both their rational and visceral selves can accept.

Chronic Over-Consumption of Unhealthy Foods

Unfortunately, the individual is just as likely to experience visceral forces tomorrow as they are today, and so is unlikely to be able to stick to the commitment they made to under-consume potato chips.

Before we analyse the implications of this, a few words on the situation the individual finds themselves in:

The individual has no say in past consumption, as that has already occurred and can't be changed. The individual can only decide on a quantity of potato chips to consume today, and make a plan for future consumption levels (which, given the individual's restraint bias, is likely to be reneged upon when the future arrives).

Without reason to believe otherwise, the individual could expect to have a total life-expectancy of around 30,000 days (~83 years). The deal the individual made with themselves concerns only two of those 30,000 days: today and tomorrow. Assuming the individual is reasonably far away from either end of those days, the quantity of past consumption prior to today, and the quantity of future consumption after tomorrow is going to be several magnitudes greater than any quantity that could possibly be consumed over the course of only today and tomorrow. This means that when a day has passed (i.e. when tomorrow becomes today), any consumption that occurred in what is now yesterday adds only a negligible amount to the total 'past' consumption. Overall, past consumption levels influence the individual in their derivation of their utility functions. In particular, parameter b (how impactful the individual judges potato chips to be on their health) is likely to depend in some way on past consumption. As total past consumption doesn't materially change from one period to the

next, the individual is unlikely to revise their parameter estimates on a daily basis, as the scale of the change in total past consumption is many times smaller than any accuracy that the individual is likely to achieve in calculating parameters (i.e. if an individual's average daily potato chip consumption in their life to date has increased by one gram, the individual's 'best-guesses' as to the impact of future consumption on their projected health trajectory are likely to remain unchanged).

So, the individual decides on consumption for today and makes plans for consumption tomorrow, in an attempt to compromise between their visceral and rational optimal consumption levels.

However, when tomorrow arrives the individual does not have the restraint they believed they would have, and is just as likely to be subject to the same visceral influences as yesterday. When the visceral influences occur, the individual feels a strong desire to consume potato chips. Unsurprisingly, this desire is just as strong as it was yesterday. The person is unable to change their past consumption, so in order to consume the amount of potato chips they'd like to, the person not only reneges on their commitment to under-consume today, but also sets a new internal deal with themselves, allowing for over-consumption today, with the plan of compensating by under-consuming tomorrow. Thus, the over consumption continues, and in every period the person actually eats:

$$x_t = \left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{2}{1+\theta^{\frac{1}{r}}}$$

The planned compensatory consumption never arrives.

Over-consumption is therefore chronic and consistent in its magnitude.

Note that it is the perceived 'wiggle room' combined with an individual's restraint bias that results in this sub-optimal consumption (as opposed to directly being the result of visceral influences). When there is no time left to procrastinate, we *do* study for that exam. When there is a personal trainer waiting for us at the gym, we *do* manage to get out of bed and do some exercise. In the world of potato chips, if there had been no opportunity for the individual to make an intertemporal deal with themselves (and/or no restraint bias that suggested they could stick with that deal), they would have overcome their visceral side and consumed rationally. Rational consumption only fails when 'wiggle room' is present.

Comparison with Present Bias Model

Optimally, a person should consume $\left(\frac{a}{b+1}\right)^{\frac{1}{r}}$ potato chips every period. However, biases cause the individual to consume sub-optimally.

In the O'Donoghue and Rabin 2006 model of present bias, per-period consumption is higher than optimal:

$$x_t^{present\ bias} = \left(\frac{a}{\beta b+1}\right)^{\frac{1}{r}}$$

In the model of internal deal-making presented here, per-period potato chip consumption is:

$$x_t^{internal\ deals} = \left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{2}{1+\theta^{\frac{1}{r}}}$$

Note that, when there is no discounting (i.e. when $\beta = \theta = 1$), both models revert back to optimal consumption levels. For all $\beta < \theta < 1$, $x_t^{internal\ deals} < x_t^{present\ bias}$, i.e. for a given level of discounting, the internal-deal model predicts lower over-consumption than the present bias model. The internal-deal model also has an upper consumption limit that is absent from the present bias results.

From a modelling perspective, in the present bias models (and other models mentioned above), the 'true' utility function is objectively correct. Therefore, when assuming that health impacts are proportional to consumption, this is equivalent to assuming that the objective science concludes that health impacts are indeed proportional to consumption. However, in the model of internal deals, both the 'rational' and 'visceral' utility functions are subjectively held by the individual. Therefore, assuming that health impacts are proportional to consumption means assuming that the individual considers the impacts to be proportional to consumption, i.e. that the individual is using a rule of thumb that equates one potato chip to b units of health harm (as opposed to a nutritionist or other expert who might have a more nuanced understanding of the relationship between consumption levels and associated health impacts).

Non-linear Health Impacts

All modelling so far has assumed that health impacts are proportional to consumption. A more general case is presented here:

Instead of proportionality, health impacts now have the form:

$$u_{1+2}^{health} = \frac{b}{n+1}x_1^{1+n} + \frac{b}{n+1}x_2^{1+n}$$

Where $n \geq 1$.

This non-linearity imposes additional restrictions on the individual, but they will still attempt to exploit what wriggle room is available.

Note that the following example removes the linearity assumption in health, and also assumes that *taste* utility is now linear. This taste assumption doesn't cause any reduction of insight/loss of generality, it just makes the maths a little easier[14].

Rational self:

$$\max_{x_1,x_2} U_{1+2}^{rational} = ax_1 + ax_2 - \frac{b}{n+1}x_1^{1+n} - \frac{b}{n+1}x_2^{1+n} + (I - x_1) + (I - x_2)$$

Optimal consumption according to rational self:

$$x_{rational} = \left(\frac{a-1}{b}\right)^{\frac{1}{n}}$$

Visceral self:

$$\max_{x_1,x_2} U_{1+2}^{visceral} = ax_1 + ax_2 - \frac{\theta b}{n+1}x_1^{1+n} - \frac{\theta b}{n+1}x_2^{1+n} + \theta(I - x_1) + \theta(I - x_2)$$

[14]An individual with non-linear taste utility AND non-linear health (dis)utility would still seek to exploit wiggle room and make an internal deal with themselves, they just might have to use an excel spreadsheet instead of algebra to figure out how many potato chips are involved in the deal.

Optimal consumption according to visceral self:

$$x_{visceral} = \left(\frac{a-\theta}{\theta b}\right)^{\frac{1}{n}}$$

As in the case of linear health impacts, when facing visceral cravings and attempting to exploit wiggle room, the individual maintains the rational level of health over two periods, while increasing today's consumption.

The rational health requirement (keeping total health the same over a two-period timeframe) is:

$$x_1^{1+n} + x_2^{1+n} = 2\left(\frac{a-1}{b}\right)^{\frac{1+n}{n}}$$

The visceral utility combined with the rational health restraint can be solved to give:

$$x_1 = \left(\frac{a-1}{b}\right)^{\frac{1}{n}} \frac{2^{\frac{1}{n}}(a-\theta)^{\frac{1}{n}}}{\left[(a-\theta)^{\frac{n+1}{n}} + (a\theta-\theta)^{\frac{n+1}{n}}\right]^{\frac{1}{n+1}}}$$

And:

$$x_2 = \left(\frac{a-1}{b}\right)^{\frac{1}{n}} \frac{2^{\frac{1}{n}}(\theta a-\theta)^{\frac{1}{n}}}{\left[(a-\theta)^{\frac{n+1}{n}} + (a\theta-\theta)^{\frac{n+1}{n}}\right]^{\frac{1}{n+1}}}$$

Which may look a little hairy, but when $\theta = 1$, this reduces to:

$$x_1 = x_2 = \left(\frac{a-1}{b}\right)^{\frac{1}{n}}$$

i.e. rational consumption

and when $\theta = 0$ this reduces to:

$$x_1 = 2^{\frac{1}{n}}\left(\frac{a-1}{b}\right)^{\frac{1}{n}}, x_2 = 0$$

When health impacts are not linear, the individual, if they could stick to their deal, would face a net loss of consumption (they can't eat as many potato chips over a two-day period as they could if they consumed optimally). For all $\theta < 1$:

$$x_1^{rational} + x_2^{rational} > x_1^{internal\ deal} + x_2^{internal\ deal}$$

But they will still attempt to exploit the wiggle room, and still fail in the attempt and continue to over consume in every period.

Price Metric of Bias

Recall that the price metric of bias is the price change (reduction) that would cause optimal (bias-free) consumption to increase to the same level as bias-induced consumption:

Explicitly labelling the price of potato chips p:

$$x_1 = \left(\frac{a}{p+b}\right)^{\frac{1}{r}} \frac{2}{1+\theta^{\frac{1}{r}}} = \left(\frac{a}{p+\gamma+b}\right)^{\frac{1}{r}}$$

This yields:

$$\gamma = (p+b)\left[\left(\frac{1+\theta^{\frac{1}{r}}}{2}\right)^{r} - 1\right]$$

In our numerical example, if we set $\theta = 0.9$ (and recalling that $b = 2, r = 0.5$) this yields a price metric of bias of$-14.6c$. This compares with a price metric of $\gamma = -20c$ in the present bias case, and again corresponds with smaller deviations from optimal in the deal-making model, and (all else being equal) smaller price changes required to price out the bias.

Naïve vs Sophisticated Consumers

A 'naïve' consumer is someone who is unaware they suffer from a bias, and so believes they will stick to the plans they've made.

It's important to remember that the bias in question is not an aspect of the visceral utility function, but rather the restraint bias that causes the individual to attempt to exploit wiggle room. As discussed earlier, when wiggle room is absent, visceral influences do not affect consumption.

The model, as set out above, therefore concerns naïve consumers: they make a deal with themselves to over-consume today on the condition they will under-consume tomorrow, and they believe themselves able to stick to their commitment. When tomorrow arrives and they renege on the deal, they are surprised at their own behaviour. Nonetheless, they continue to set deals with themselves, overconsume, and be surprised when the deal is not fulfilled.

A 'sophisticated' consumer, on the other hand, knows they suffer from a bias that will cause them to be unable to stick to the plans they've made. This effectively removes any wriggle room, forcing the consumer to ignore their visceral self and consume rationally.

Effectively, knowing you suffer from restraint bias is sufficient to eliminate the bias.

Longer Commitment Periods

In the model presented above, the consumer, whether naïve of sophisticated, is always planning over two periods: today and tomorrow. This has somewhat of a real-world psychological base, as our internal justifications tend to take the form: "It's ok if I over-consume today because I'll eat well/start exercising/quit drinking (etc) starting from tomorrow."

But when we make such a statement, what do we really mean? Do we mean that we will overly restrict our eating tomorrow only, after which we presumably intend to return to rational, moderate consumption? Do we mean that we intend to overly restrict our potato chip consumption forever? Quite often, it seems we mean we will restrict ourselves until we meet some target date/health level (i.e. we take a rather nebulous approach to the definition of the word 'tomorrow'), after which we plan to set our eating at the rational optimum level.

The two-period model set out above can be considered as a simple 'reduced form' where the more realistic variant is that people make multi-day commitments, but random fluctuations in exposure to visceral influences can 're-set' the commitment period, and cause the person to

over-consume via a new internal-deal. However, in spite of an individual's ability to stick with a deal on some days, total consumption will still be excessive.

If exposure to visceral influences triggers cravings and the reneging on any deal, an individual will be able to stick to a deal/consume optimally only if they are not exposed to visceral influences. If exposure on a given day is not guaranteed, then an individual will stick to any deal they set for themselves in the previous period. The following table illustrates four different combinations (of exposure or non-exposure to visceral influences) over a two-day period:

Table 1: Consumption consequences of exposure and non-exposure to visceral influences

Today		**Tomorrow**		**Total Consumption**
Visceral Influences	**Consumption**	**Visceral Influences**	**Consumption**	
Exposed	$\left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{2}{1+\theta^{\frac{1}{r}}}$	*Exposed*	$\left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{2}{1+\theta^{\frac{1}{r}}}$	$\left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{4}{1+\theta^{\frac{1}{r}}}$
Exposed	$\left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{2}{1+\theta^{\frac{1}{r}}}$	*Not exposed*	$\left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{2\theta^{\frac{1}{r}}}{1+\theta^{\frac{1}{r}}}$	$2\left(\frac{a}{1+b}\right)^{\frac{1}{r}}$
Not exposed	$\left(\frac{a}{1+b}\right)^{\frac{1}{r}}$	*Exposed*	$\left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{2}{1+\theta^{\frac{1}{r}}}$	$\left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{3}{1+\theta^{\frac{1}{r}}}$
Not exposed	$\left(\frac{a}{1+b}\right)^{\frac{1}{r}}$	*Not exposed*	$\left(\frac{a}{1+b}\right)^{\frac{1}{r}}$	$2\left(\frac{a}{1+b}\right)^{\frac{1}{r}}$

On any given two-day period (which could correspond to any combination of exposure and non-exposure to visceral influences) an individual will *at best* consume optimally, and on average will over-consume.

From a modelling perspective, it is possible to either allow for a more nebulous definition of tomorrow (the individual considers 'tomorrow' to be the next three months, for example), or equally the number of periods in the commitment can be explicitly modelled:

In the case of K periods (making a deal between today and $K-1$ periods in the future), the visceral utility is:

$$\max_{x_1 \dots x_k} U_{1-K}^{visceral} = \frac{a}{1-r} x_1^{1-r} + \theta(I - x_1) - \theta b x_1 + \sum_{t=2}^{K} \theta \frac{a}{1-r} x_t^{1-r} - \theta b x_t + \theta(I - x_t)$$

Subject to the restraint:

$$\sum_{t=1}^{K} x_t = K\left(\frac{a}{1+b}\right)^{\frac{1}{r}}$$

This yields:

$$x_1 = \left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{K}{1+(K-1)\theta^{\frac{1}{r}}}$$

And:

$$x_2,..,x_k = \left(\frac{a}{1+b}\right)^{\frac{1}{r}} \frac{K\theta^{\frac{1}{r}}}{1+(K-1)\theta^{\frac{1}{r}}}$$

The larger the number of periods the individual is going to take to 'make up' for today's over-consumption, the greater today's over consumption can be.

Equally, the larger the number of periods the individual is going to take to 'make up' for today's over-consumption, the closer planned future periods can be to rational consumption levels.

Unfortunately, this increase of future planned consumption to near rational levels is no help to the naïve consumer, as the naïve consumer will consistently renege on the deal and overconsume. Since today's over-consumption is higher when the internal deal covers a greater number of periods, naïve consumers are worse off when compared with only trading-off between today and tomorrow.

Sophisticated consumers will not make a deal over any commitment period.

Chapter 4: Discussion

The model of internal deals presented in this dissertation has at its core *restraint bias* and the attempted exploitation of perceived flexibility to achieve both rational health and visceral taste outcomes. Unlike models of present bias, this model has no opinion as to the relative value of the rational and visceral utility functions, other than the relative values placed on them by the individual.

The bias that needs to be 'corrected' in the model of internal deals is not related to the visceral utility function, but is restraint bias. Restraint bias can be 'solved' by removing flexibility in achieving optimal health outcomes.

Real-World Applicability

The model of internal deals presented here has a few key assumptions. The first is that we have a rational self that knows we should eat unhealthy food only in moderation. This assumption matches with individual lived experiences: even in the midst of a binge, we know that we shouldn't be eating that way. What we consider to be 'moderate' may differ from person to person, and a person living with type 2 diabetes may have more stringent requirements than a person who is young, fit and has never experienced serious ill-health, but we each have a feel for what an appropriate level of consumption should be.

And yet, if we're someone who craves potato chips, or whether we prefer chocolate, alcohol or cigarettes, we often find ourselves consuming too much of that good, in spite of our best intentions. Furthermore, it seems we often exhibit naïve behaviour: We regularly commit to 'eating well starting from tomorrow', and regularly abandon such a commitment when tomorrow arrives. Given the extent of the obesity epidemic, it seems most of us are naïve, most of the time.

Objectively, it seems unlikely that an individual would repeatedly make a deal with themselves, fail to stick with it, and then remake the same deal without ever changing their behaviour. However, in reality humans often do exactly that; repeatedly re-joining a gym despite past evidence that they won't attend, or committing to a diet despite past failures to stick with a diet. Therefore, this naïve behaviour that leads to consistent over-consumption seems to have a strong grounding in reality.

What is interesting, is that naïve and sophisticated behaviour seem to manifest themselves in the same individual (although not simultaneously). This means there are times in our lives when we over consume by making deals we falsely believe we can keep, and times when we manage to consume only in moderation. This moderate consumption could arise because we are sometimes not exposed to visceral influences, or because we sometimes can exhibit sophisticated behaviour.

How can we alternate between both naïve and sophisticated behaviours? Within our individual life experiences, we have evidence we can draw on that encourages both naïve and sophisticated thinking: there have been occasions in our life where we *have* been able to show restraint and stick to a commitment, and our naïve self can draw on these examples when making an internal deal and believing it can commit to that deal. Equally, there have been occasions when we have succumbed to temptation despite yesterday's promise not to, and our sophisticated selves can draw on these memories when determining that an internal deal is not possible and we must consume optimally.

The number of paths towards optimal health is a larger concern in the real world than is manifested in the model of internal deals. In the model, the individual only has the option of

shifting part of tomorrow's consumption to today. In the real world, there are near-infinite possibilities to achieve optimal health. As well as under-consuming the good in question tomorrow, the individual could also exercise (through weights, cardio, yoga, walking, running…), they could do the exercise now, or in an hour, or in a week, they could do a little, often, or a marathon in a week, they could build muscle as a way to boost their resting metabolism, they could give up some other food, or go on a keto diet, a paleo diet… and it is precisely the large range of possibilities to achieve good health that makes good health so hard to achieve. When so many internal deals are possible, the individual has too much wriggle room to ever hope to consume optimally.

Comparison with Other Visceral Models

The model proposed by Bernheim and Rangel (2004) assumed that people subject to visceral forces would always consume, and that the visceral force overwhelmed any rational decision making.

The model of internal deals, on the other hand, does not propose that visceral influences are all-powerful, but rather that the individual can be simultaneously aware of both their visceral cravings and their rational knowledge about appropriate consumption, and needs to make an appropriate decision based on these competing factors. When examining the real world, it seems we are aware of how much we should be consuming even when bingeing, and we consume how much we should when we feel we have no other option.

According to the model by Gul and Pesendorfer (2001), people overcome visceral forces using costly will-power. Will-power is not a factor in the model of internal-deals. If wiggle room exists, a consumer (or at least a naïve consumer) will attempt to exploit it. If no wiggle room exists, the individual will consume optimally.

Policy Implications

Like models of present bias, the model of internal deal-making suggests that taxation is one method to return individuals to optimal consumption levels. When consumers exhibit naïve over-consumption, the optimal correctional tax is positive. The price-metric of bias gives an indication of the size of the correction needed to correct the bias. Taxes are only one type of policy instrument, however, and other non-tax policies (either in addition to or instead of taxes) could be more appropriate.

Policies that reduce exposure to visceral influences (banning advertising, shutting down Wi-Fi after 10pm so people go to bed early and sleep well, implementing a 4-day work week so people aren't so stressed) could also be beneficial.

However, in internal deal-making, the root cause of the bias is the flexibility in the number of consumption paths that achieve optimality. Policies that reduce consumption choices could include "next day delivery", where junk food purchases are only made available to the consumer the following day (or where a discount is available if delivery is delayed). Since internal deals are often used to justify impulsive consumption desires, restricting consumption to the day following the purchase would reduce and perhaps eliminate the occurrence of such deals. Other policies that either reduce the number of paths towards optimality or reduce the making of internal-deals as a result of path multiplicity would also be effective in allowing people to consume optimally.

References

Allcott, H., Lockwood, B. & Taubinsky, D. (2019). Regressive Sin Taxes, With an Application to the Optimal Soda Tax. *The Quarterly Journal of Economics*, 1557-1626

Ariely, D. & Wertenbroch, K. (2002). Procrastination, Deadlines, and Performance: Self-Control by Precommitment. Psychological Science, 13(3), 219-224

Benzion, U., Rapoport, A. & Yagil, J. (1989) Discount Rates Inferred from Decisions: An Experimental Study. Management Science, 35(3), 270-284

Bernheim, B. (2016). The good, the bad, and the ugly: a unified approach to behavioral welfare economics. Journal of Benefit-Cost Analysis, 7(1), 12-68

Bernheim, B. & Rangel, A. (2004). Addiction and cue-triggered decision processes. *American Economic Review*, 94(5), 1558–1590.

Bernheim, B. & Rangel, A. (2007a). Toward Choice-Theoretic Foundations for Behavioral Welfare Economics. *The American Economic Review, 97*(2), 464-470

Bernheim, B. & Rangel, A. (2007b). Behavioral public economics: welfare and policy analysis with fallible decision-makers. In: P. Diamond and H. Vartianen, (Eds.), *Behavioral Economics and Its Applications*, (pp. 7-77) Princeton University Press.

Bernheim, B.D. & Taubinsky, D. (2018). Behavioural Public Economics. In B. Bernheim, S. Dellavigna & D. Laibson (Eds.). *Handbook of Behavioural Economics: Foundations and Applications* (pp. 381-488). Amsterdam : North-Holland: Elsevier Science & Technology.

Bollinger, B., Leslie, P. & Sorensen, A. (2011). Calorie posting in chain restaurants. *American Economic Journal: Economic Policy,* 3 (1), 91–128.

Briggs, A., Mytton, O., Kehlbacher, A., Tiffin, R., Rayner, M. & Scarborough, P. (2013). Overall and income specific effect on prevalence of overweight and obesity of 20% sugar sweetened drink tax in UK: econometric and comparative risk assessment modelling study, *BMJ,* 347:f6189

Camerer, C., Babcock, L., Loewenstein, G. & Thaler, R. (1997). Labor Supply of New York City Cabdrivers: One Day at a Time. *Quarterly Journal of Economics*, 112(2), 407-441

Chetty, R., Looney, A. & Kroft, K., (2009). Salience and taxation: theory and evidence, *American Economic Review,* 99 (4), 1145–1177.

DellaVigna, S. & Malmendier, U., (2006). Paying Not to Go to the Gym, *American Economic Review*, 96 (3), 694-719

Dharmasena, S. & Capps, O. (2012). Intended and unintended consequences of a proposed national tax on sugar sweetened beverages to combat the US obesity problem. *Health Economics,* 21, 669–694.

Dimitri N. & van Eijck J. (2012) Time Discounting and Time Consistency. In: J. van Eijck and R. Verbrugge, (eds) *Games, Actions and Social Software. Lecture Notes in Computer Science*, *vol 7010*. Springer, Berlin, Heidelberg.

Dodd,M. (2008). Obesity and time-inconsistent preferences. *Obesity Research & Clinical Practice, 2*(2), 83-89

Farhi, E. & Gabaix, X. (2020). Optimal Taxation with Behavioural Agents. *American Economic Review, 110*(1), 298-336

Frederick, S., Loewenstein, G. and O'Donoghue, T. (2002). Time discounting and time preference: a critical review. *Journal of Economic Literature,* 40 (2), 351–401.

Gostin, L. & Gostin, K. (2009). A broader liberty: J.S. Mill, paternalism and the public's health. *Public Health*, 123(3), 214-221

Gruber, J. and Köszegi, B. (2001). Is addiction "rational"? Theory and evidence. *The Quarterly Journal of Economics*, 116 (4), 1261–1303.

Gul, F. & Pesendorfer, W. (2001). Temptation and self-control. *Econometrica,* 69 (6), 1403–1435

Harkanen, T., Kotakorpi, K., Pietinen, P., Pirttila, J., Reinivuo, H. & Suoniemi, I. (2014). The welfare effects of health-based food tax policy. *Food Policy, 49*, 196-206

Hoch, S. & Loewenstein, G. (1991). Time-inconsistent Preferences and Consumer Self-Control. *Journal of Consumer Research*, 17(4) 492-507

Kahneman, D. (1994). New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics,* 150 18–36.

Köszegi, B. & Szeidl, A. (2013). A Model of Focusing in Economic Choice. *Quarterly Journal of Economics,* 128(1), p53-104

Kotakorpi, K. (2008). The incidence of sin taxes. *Economic Letters, 98*(1), 95-99

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics,* 112 (2), 443–478.

Lockwood, B. & Taubinsky, D. (2017). Regressive Sin Taxes. *National Bureau of Economic Research,* Working Paper 23085

Loewenstein, G. (1996). *Out of Control: Visceral Influences on Behavior*. Organizational Behavior and Human Decision Processes, 65(3), 272-292

Loewenstein, G., O'Donoghue, T. & Rabin, M. (2013). Projection Bias in Predicting Future Utility. *The Quarterly Journal of Economics, 118(4):1209-1248*.

Mandel, N., Scott, M., Kim, S. and Sinha, R. (2017). Strategies for improving self-control among naïve, sophisticated, and time-consistent consumers. Journal of Economic Psychology, 60, 109-126

Nordgren, L., van Harreveld, F. & van der Pligt, J. (2009). The restraint Bias: How the Illusion of Self-Restraint Promotes Impulsive Behavior. *Psychological Science*, 20(12):1523-1528

O'Donoghue, T. & Rabin, M. (1999). Doing it Now or Later. *American Economic Review, 89(1):103-124*

O'Donoghue, T. & Rabin, M. (2003). Studying Optimal Paternalism, Illustrated by a model of Sin Taxes. *American Economic Review, 93(2):186-191*

O'Donoghue, T. & Rabin, M. (2006). Optimal Sin Taxes. *Journal of Public Economics, 90, 1825-1849*

O'Donoghue, T. & Rabin, M. (2015). Present Bias: Lessons Learned and to Be Learned. *American Economic Review, 105(5):273-279*

Thaler, R.H. and Sunstein, C.R. (2003). *Libertarian Paternalism,* The American Economic Review, 93(2) 175-179.

Obesity and overweight (Factsheet). (2020, April 1). Retrieved July 15, 2020, from https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Obesity statistics. (n.d). Retrieved July 16, 2020, from https://www.health.govt.nz/nz-health-statistics/health-statistics-and-data-sets/obesity-statistics

Appendix

Recall:

$$U_{1+2}^{visceral} = \frac{a}{1-r}x_1^{1-r} + \theta\frac{a}{1-r}x_2^{1-r} - \theta b(x_1 + x_2) + \theta(I - x_1) + \theta(I - x_2)$$

The individual seeks to optimise this utility, subject to the constraint:

$$x_1 + x_2 = 2\left(\frac{a}{1+b}\right)^{\frac{1}{r}}$$

$2\left(\frac{a}{1+b}\right)^{\frac{1}{r}}$ is a constant. Therefore, for ease of calculation, it can (temporarily) be labelled as c. i.e.:

$$x_1 + x_2 = c$$

This can be solved using the Lagrange method.

The Lagrangian:

$$L = \frac{a}{1-r}x_1^{1-r} + \theta\frac{a}{1-r}x_2^{1-r} - \theta b(x_1 + x_2) + \theta(I - x_1) + \theta(I - x_2) - \lambda(c - x_1 - x_2)$$

W.r.t x_1:

$$\frac{\partial L}{\partial x_1} = ax_1^{-r} - \theta b - \theta + \lambda = 0$$

$$x_1^{-r} = \frac{\theta b + \theta - \lambda}{a}$$

$$x_1 = \left(\frac{a}{\theta b + \theta - \lambda}\right)^{\frac{1}{r}}$$

W.r.t x_2:

$$\frac{\partial L}{\partial x_2} = \theta a x_2^{-r} - \theta b - \theta + \lambda = 0$$

$$x_2^{-r} = \frac{\theta b + \theta - \lambda}{\theta a}$$

$$x_2 = \left(\frac{\theta a}{\theta b + \theta - \lambda}\right)^{\frac{1}{r}}$$

W.r.t λ:

$$\frac{\partial L}{\partial \lambda} = x_1 + x_2 - c = 0$$

$$x_1 + x_2 = c$$

Solving for λ:

$$c = \left(\frac{a}{\theta b + \theta - \lambda}\right)^{\frac{1}{r}} + \left(\frac{\theta a}{\theta b + \theta - \lambda}\right)^{\frac{1}{r}}$$

$$c = \left(\frac{1}{\theta b + \theta - \lambda}\right)^{\frac{1}{r}} \left(a^{\frac{1}{r}} + (\theta a)^{\frac{1}{r}}\right)$$

$$c^r = \frac{1}{\theta b + \theta - \lambda} \left(a^{\frac{1}{r}} + (\theta a)^{\frac{1}{r}}\right)^r$$

$$\theta b + \theta - \lambda = \left(\frac{a^{\frac{1}{r}} + (\theta a)^{\frac{1}{r}}}{c}\right)^r$$

$$\lambda = \theta b + \theta - \left(\frac{a^{\frac{1}{r}} + (\theta a)^{\frac{1}{r}}}{c}\right)^r$$

Solving for x_1:

$$x_1 = \left(\frac{a}{\theta b + \theta - \lambda}\right)^{\frac{1}{r}}$$

$$x_1 = \left(\frac{a}{\theta b + \theta - \left(\theta b + \theta - \left(\frac{a^{\frac{1}{r}} + (\theta a)^{\frac{1}{r}}}{c}\right)^r\right)}\right)^{\frac{1}{r}}$$

$$x_1 = \left(\frac{a}{\left(\frac{a^{\frac{1}{r}} + (\theta a)^{\frac{1}{r}}}{c}\right)^r}\right)^{\frac{1}{r}}$$

$$x_1 = c\frac{a^{\frac{1}{r}}}{a^{\frac{1}{r}} + (\theta a)^{\frac{1}{r}}}$$

$$x_1 = \frac{c}{1 + \theta^{\frac{1}{r}}}$$

Solving for x_2:

$$x_2 = \left(\frac{\theta a}{\theta b + \theta - \lambda}\right)^{\frac{1}{r}}$$

$$x_2 = \left(\frac{\theta a}{\theta b + \theta - \left(\theta b + \theta - \left(\frac{a^{\frac{1}{r}} + (\theta a)^{\frac{1}{r}}}{c}\right)^r\right)}\right)^{\frac{1}{r}}$$

$$x_2 = \left(\frac{\theta a}{\left(\frac{a^{\frac{1}{r}} + (\theta a)^{\frac{1}{r}}}{c} \right)^r} \right)^{\frac{1}{r}}$$

$$x_2 = c \frac{\theta^{\frac{1}{r}} a^{\frac{1}{r}}}{a^{\frac{1}{r}} + (\theta a)^{\frac{1}{r}}}$$

$$x_2 = \frac{\theta^{\frac{1}{r}} c}{1 + \theta^{\frac{1}{r}}}$$

Returning c to its original value gives:

$$x_1 = \left(\frac{a}{1+b} \right)^{\frac{1}{r}} \frac{2}{1 + \theta^{\frac{1}{r}}}$$

And:

$$x_2 = \left(\frac{a}{1+b} \right)^{\frac{1}{r}} \frac{2\theta^{\frac{1}{r}}}{1 + \theta^{\frac{1}{r}}}$$