**Economics Working Paper Series**

**2019/003**

# A Review of Economic Dual-Self Modelling

Iain P. Embrey

The Department of Economics
Lancaster University Management School
Lancaster LA1 4YX
UK

LUMS home page: http://www.lancaster.ac.uk/lums/

# A Review of Economic Dual-Self Modelling

**Abstract**

   Dual-self decision theory generalises the canonical economic model by admitting multiple, possibly conflicting, decision criteria. A typical dual-self model will formalise psychological or neuroscientific descriptions of the human decision-making process into an economic model of that process, and apply that model to provide an unified explanation for several behavioural anomalies. In this paper, we compare the foundations of Neoclassical and dual-self decision theories, we develop a generalised decision framework that nests both decision theories, and we use that framework to provide a comprehensive taxonomy of the economic dual-self literature. We also discuss the relative merits of each existing dual-self approach, and suggest avenues for future research.

**Keywords:** Dual-Self, Dual Self, Decision Theory, Behavioural

**JEL Codes:** D01, D91, D81, B41

# 1   Introduction

The first economic dual-self model was developed by Thaler and Shefrin, and published in 1981. Their insight was that several aspects of consumer behaviour could be explained by attributing to each individual two distinct entities: a 'planner' who seeks to maximise lifetime utility, and a 'doer' who seeks to maximise present period utility. Those dual selves embody the same quintessential human conflict between deliberative and impulsive motivations that underlies much of the current dual-self literature, although there is now substantial variation in the operationalisation of that conflict. In this paper we survey the assumptions, the applications, and the implications from a large portion of the economic dual-self literature, and in doing so we extend substantially the existing selective review of Brocas & Carrillo (2014$a$).

   Our review identifies five distinct modelling approaches within the literature. Accordingly, we taxonomise existing studies into Tables 1, 2a,b, 3 and 4, in order to analyse the achievements, the limitations, and the future potential of each approach in Section 3.2. Our analysis exposes a distinct disparity between the implications of those various

modelling approaches, however it also enables us to identify a single, generalised, decision framework that has the potential to bring together many of those disparate strands of the literature. Our generalised decision framework: nests the traditional 'single-self' paradigm; encompasses the majority of existing dual-self models as special cases; and is also closely aligned with the psychological and neuroscientific literatures as distilled by Evans & Stanovich (2013) and Bechara (2005). We present the generalised framework in Section 3.1, and we conclude that it provides a promising direction for future research.

Traditional Expected Utility Theory is founded upon the principle that all outcomes must admit one unique utility valuation. In contrast, our generalised decision framework allows utility to be vector-valued. Section 2 compares the foundational assumptions and implications of traditional Expected Utility Theory with those of our generalisation. In doing so we determine a priori that the more mathematically elegant 'single-self' approach should be favoured in situations characterised by a single clear objective such as profit maximisation, but that the more general dual-self approach should be favoured in situations characterised by multiple, conflicting, motivations. Our discussion in Section 2 is reminiscent of those in classic texts such as Marshall's (1890) Principles of Economics, except that, where Marshall concluded that economic analysis can only be applied in circumstances with monetarily quantifiable motivations, we conclude that circumstances with diverse behavioural motivations could become tractable under our generalised decision framework. For the reader who wishes to focus on existing implementations of dual-self theory, Section 2 may be passed over: A holistic summary and conclusion is provided in Section 4.

## 2    The Foundations of Dual-Self Theory

The dual-self paradigm for human decision-making is well-established in the psychological literature, widely supported by neuroscience, and increasingly recognised as a valuable approach to economic modelling[1]. Where Neoclassical models assume a single functional form for decision utility, dual-self models specify multiple utility functions each of which operationalises one possible *decision criterion*, or way-of-thinking. That generalisation provides a descriptive theory of human action under multiple motivations, without requiring those motivations to be directly comparable. This section will outline the strengths and limitations of Neoclassical economic modelling, and explain the ways in which its dual-self generalisation can overcome some of those limitations. A thorough discussion of the restrictions required by the Neoclassical approach can be found in the opening chapters of either Marshall (1890) or Von Mises (1949).

The Neoclassical decision framework provides both an elegant normative theory and an intuitively appealing descriptive theory of human decision-making – for any situation which is naturally characterised by a single objective function. Traditional economic

problems, such as profit maximisation, are therefore well served by the canonical theory. However, modern behavioural economics applies that same theory to situations characterised by multiple, conflicting, motivations. Such applications of expected utility theory imply that those diverse motivations can be completely represented by their image under some hypothetical projection onto a single utility dimension, yet that cardinal conceptualisation of utility has been widely discredited for over a century.

Cardinal Utilitarianism, as proposed by Bentham (1789), declares that individuals should, and often do, act deliberately to maximise their predicted utility. Conversely, modern utility theory is founded on the observation that, since individuals have preferences, they act *as though they were* maximising some decision utility function. That observation was formalised by Samuelson (1938) under the revealed preference paradigm, and derived by von Neumann & Morgenstern (1947) under their axiomatisation of Expected Utility Theory. A concise statement of their result is that every outcome must admit a well-defined decision utility valuation. Where Bentham postulated this foundational result[2], von Neumann-Morgenstern, Savage (1954), and others have derived it from more primitive axioms. Nevertheless, any such axiomatisation necessarily requires human action to manifest complete, transitive, and consistent preferences – an assumption set which is frequently contradicted by observed behaviour (Rabin 2002).

The Neoclassical decision framework therefore faces challenges to both its intuitive and its theoretical validity in non-catallactic situations. Nevertheless, the critical limitation of that framework is that it can only describe a somewhat particular decision-making mechanism: Homo economicus always weighs up her various motivations through concious deliberation. Homo sapiens frequently do not (Smith 1759)[3]. Furthermore, whenever agents are modelled *as though they were* utility maximisers, their decision-making processes are being described as a black box. Thus the canonical model provides a descriptive theory of only the outcome from behavioural decisions, whilst it is the decision mechanism which is of central interest in the design of policy or interventions.

Despite its limitations, there are at least three key applications for which the Neoclassical decision framework is ideally suited. First, by prescribing a unique functional form as its representative agent's objective, the canonical model identifies its agent's normatively optimal response to any accurately operationalised situation. This application of economic theory is precisely the science of Praxeology proposed by Von Mises (1949). Second, the canonical model aptly describes the process of deliberative decision-making, in circumstances where "the advantages and disadvantages of any particular action" may be "reckoned-up". This is precisely the restricted domain for which Marshall (1890, p.17) expounded his Principles of Economics. Third, through ad-hoc adaptations of the utility function, the canonical theory can provide tractable *as if* characterisations of many non-pecuniary motivations. Such *as if* characterisations provide remarkably accurate predictions of behavioural outcomes for particular decision situations (Crawford, Costa-

Gomes & Iriberri 2013, Kosfeld et al. 2017).

The novelty of dual-self theory is that it explicitly models the interaction between conflicting decision criteria. That approach can produce an unified explanation of diverse behavioural anomalies, and can simultaneously suggest mechanisms for the processes of human decision-making. Section 3 assesses the contributions of existing dual-self theory toward each of these objectives.

Progress toward these objectives has been substantial. For example Fudenberg & Levine (2006) provide an unified explanation for present-bias, the effect of orthogonal cognitive load on decision outcomes, and the paradox of risk aversion in the large and the small (ex Rabin 2000). Additional behavioural anomalies, such as the effects of framing, priming, habit, and one's current emotional state can equally be understood if they cue certain decision criteria (Bernheim & Rangel 2004), or if they modulate the strength of impulsive motivations (Loewenstein, O'Donoghue & Bhatia 2015). Furthermore, the foundational principle of dual-self models, that every outcome must admit a well-defined utility valuation *in each of two dimensions*, is, in a precise sense, the weakest generalisation of the canonical foundation that can predict temporal inconsistency (which in turn predicts intransitivities).

The power of dual-self theories to provide unified explanations for behavioural anomalies is derived from the hypothesis that each anomaly may represent a special case of some single, fundamental, anomaly of human behaviour: specifically the tendency of individuals to act against their own long-term self-interest (Baumeister 2003). That hypothesis suggests that each such anomaly may be rooted in the quintessential human conflict between deliberative and impulsive motivations; a conflict which is naturally operationalised by dual-self theory. Indeed, almost the whole literature synthesised in Tables 1-4 models that same quintessential conflict, although it is described in various alternative terminologies[4]. The various, approximately synonymous, terms adopted by those papers reflect the idiosyncrasies of their respective models, which are frequently motivated by one particular psychological, experimental, or neuroscientific characterisation of decision-making.

This paper will not reproduce the well-established evidence base for the dual-self paradigm; instead the reader is referred to several prominent papers which have exposed the pervasive importance of admitting multiple, possibly conflicting, decision criteria in order to understand observed behaviours. Loewenstein (1996) provides an excellent general discussion and characterisation of the visceral impulses which affect human decision-making, Cohen (2005) describes the neuroscientific and evolutionary validation of that characterisation, and Schelling (1984) discusses its implications for optimal self-control behaviour in diverse situations.

Most existing dual-self models focus on a rational agent's optimal response to the quintessential human conflict, under the assumption that deliberative selves have complete information regarding both the existence of, and the optimisation problem of, each

alternative decision criterion. Whilst that assumption provides valuable insights, it is "clearly unrealistic" (Brocas & Carrillo 2003, p.xviii). That assumption also implicitly maintains the Neoclassical requirement that the population can be adequately described by the singular utility formulation of some hypothetical representative agent. One alternative assumption would be to specify that each individual's deliberative self will override their impulsive self with a probability drawn from some individual- and situation-specific distribution. This alternative assumption forms the basis of the generalised decision theory set out in Section 3.1, which nest the models of Tables 1-4 as progressively less restrictive special cases. The final model listed in Table 4 (Embrey 2017) applies the fully generalised decision theory to provide an elementary understanding of the human life-course, under which divergent human-capital outcomes result from heterogeneity in ways-of-thinking.

The novelty of the generalised decision theory is that it describes a process of human decision-making. That descriptive theory could be reduced to a normative theory by stipulating which singular decision criterion its agents *should* apply. As an illustrative example, consider the everyday decisions which determine individuals' educational, employment, and health outcomes – for example whether to: do homework, attend class, actively seek employment, exercise or watch t.v., cook or order a take-away,...[5] These decisions manifest the quintessential human conflict between deliberative and impulsive reasoning, wherein a normative theory is provided by the imposition of some functional form by which future consequences should be traded off against present desires. Note that the implementation of such a trade-off itself requires deliberative thinking, whence Neoclassical decision theory cannot describe heterogeneity in thought processes.

The canonical theory therefore provides three candidate mechanisms by which individuals could make normatively poor decisions: i. their estimated likelihood of future consequences may be too low, ii. their estimated payoffs from those future consequences may be too small, or iii. those future payoffs may be too heavily discounted. Under the generalised model there is an additional candidate mechanism: iv. individuals may have a positive probability of acting without evaluating the future consequences of their action. The distinction between mechanisms iii. and iv. is substantive. The former implies that, except for chance occurrences, individuals deliberatively choose their own socio-economic outcomes, whereas the latter implies that those outcomes are an unintended consequence of individuals' socially-determined ways-of-thinking. The latter hypothesis has been firmly adopted in the health inequalities literature, where unhealthy decisions are considered to be a product of socio-economic determinants rather than individual choice (Graham 2007, Watt 2007).

The generalised decision theory therefore provides significant new insights for any situation where the probability of unconsidered action is non-negligible. A wide class of situations are likely to satisfy that condition: certainly every person has, on occa-

sion, acted without first considering the consequences of that action, and it may even be that such occasions outnumber their complement. Furthermore, there are particular situations, such as youth, intoxication, addiction, sleep deprivation, malnutrition, stress, poverty, and morbidity, under which unconsidered decision-making is particularly likely (Donohew et al. 2000, Goldman 2012, Mani et al. 2013). By explicitly modelling heterogeneous ways-of-thinking, both within and between individuals, the generalised model provides a rich descriptive theory of such situations. Moreover, since the impulsive decision criterion predicts purely subconscious responses, the proposed theory applies equally to situations which the agent need not even identify as a decision problem.
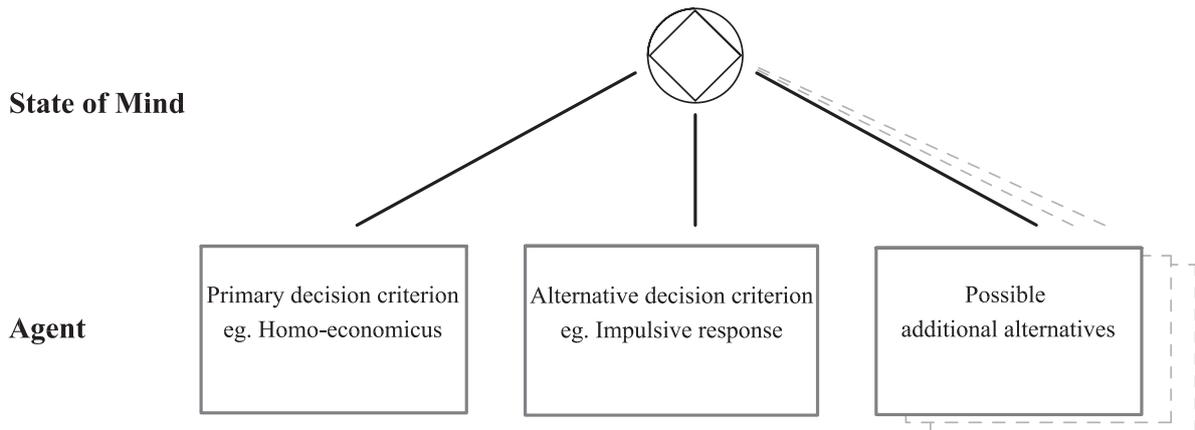
Under Neoclassical decision-theory, individual differences arise as consequences of heterogeneity in tastes. The generalised model supplements this understanding by additionally admitting heterogeneity in thought-processes. That concept has support from authors in psychology, economics, and decision theory, who have long described human action as the result of a conflict between decision criteria (e.g. Edwards 1954, Ellsberg 1961, Raiffa 1969). Recent advances in Neuroscience have also been remarkably consistent with this description (Bechara 2005). Furthermore, one might observe that the tendency of an individual to adopt certain decision criteria manifests their levels of conscientiousness, extraversion, agreeableness, or openness to experience - thus it underpins four of psychology's big five personality traits. Finally I suggest that the fifth trait, neuroticism, could be characterised by a lack of consistency and consciousness an individual's determination as to which decision criterion should be acted upon.

## 3    The Dual-Self Literature

### 3.1    A Generalised Theory of Decision-Making

In order to categorise and compare dual-self modelling approaches, it will be helpful to first outline a generalised decision framework which nests both dual-self and Neoclassical models of decision-making. Figure 1 presents this framework as a dynamic game in extensive form. Each box in Figure 1 represents one possible 'self', that is one possible decision criterion that could be enacted by the agent. In particular, traditional economic models include just one decision criterion: typically an expected-utility representation of profit maximisation. Unsurprisingly, dual-self models typically incorporate two possible decision criteria, although, in principle, the generalised behavioural framework could accommodate any number of alternative selves.[6]

In itself, it is not novel for economic theory to incorporate utility components that represent distinct motivational dimensions. This is common practice in behavioural economics. However, it is also common practice to splice those distinct utility components into a single decision criterion, which requires the imposition of some functional form

Figure 1: *A Generalised Decision Framework*



by which those distinct utility dimensions should be traded-off. Within the generalised framework of Figure 1, such practice assumes that the first stage of an agent's decision can be described by some meta-rational expected utility process. We exposed the restrictiveness of that assumption in Section 2, but it is nevertheless commonly maintained outside of the dual-self literature; typically with neither justification nor acknowledgement. The novel contribution of dual-self theory is therefore to explicitly consider how best to model the interaction between conflicting decision criteria. We refer to outcome of that interaction as the generalised decision-maker's *state of mind*.

Economic dual-self models use one of four approaches to determine their decision-maker's state of mind. The models summarised in Table 1 consider that an agent's state of mind is determined by the type of decision that she is faced with. An archetype of this approach is the planner-doer paradigm of Thaler & Shefrin (1981). The models of Table 2 assume an expected-utility meta-rationality, but they do so explicitly and based upon some specific psychological or neuroscientific description of decision-making. In contrast, the models of Table 3 adopt a game-theoretic interpretation of the interaction between alternative decision criteria, by assuming that each self has perfect but incomplete information. Finally, the models of Table 4 replace any strict assumption over the agent's state of mind with a conditional probability distribution over the set of possible decision criteria. Since those models consider situations with binary outcomes, this can be done without loss of generality because one or other decision-criterion will necessarily prevail in such situations.

Although the behavioural literature introduces a great many non-standard motivations, most dual-self theories describe some instance of the same quintessential human conflict between reasoned and instinctive actions. Furthermore, most dual-self theories encode each of those decision criteria into an utility maximisation problem. Nevertheless, there is no requirement for each decision criterion to implement Expected Utility

Theory; for example Bernheim & Rangel (2004) specify a degenerate 'hot' decision crite-rion, wherein an addict will always consume substances. Future research could therefore seek to incorporate other non-standard decision criteria, such as the Case-Based Decision Theory of Gilboa & Schmeidler (1995), into the generalised decision framework. Since there is little inter-model variation in decision criteria, we now structure our taxonomy and review of the dual-self literature around its alternative approaches toward modelling the decision-maker's state of mind.

### 3.2   A Review of the Economic Dual-self Literature

This section provides a taxonomic review of the economic dual-self literature. Tables 1-4 categorise each model according to its description of the human decision-making process, outline the situations to which it is applied, and summarise the main behavioural insights which result from those applications. An exhaustive discussion of each existing model is not provided, however the key achievements and limitations of each modelling approach are highlighted, and potential future developments are discussed.

Dual-self models are generally conceived in order provide theoretical explanations for behaviours which deviate from the predictions of standard economic theory. The experimental literature has documented many such behaviours, which range from the incontrovertibly flawed decision-making that arises from misconceptions and misrepre-sentations, to decisions which seem excessively present-biased, risk-averse, or visceral when compared with some exogenously defined normative standard. The dual-self liter-ature addresses each class of normatively suboptimal decision-making through separate approaches to modelling the decision-maker's state of mind.

The models of Table 3 address incontrovertibly flawed decision-making. These models are expressly developed to explore the cognitive implications of imperfect communication between neurological systems, and they do so by modelling a co-ordination or adversar-ial game between selves who each possess private information. It is a strength of this approach that, not only is imperfect information processing a neuroscientific certainty (Brocas 2012), but its contradiction is intuitively unjustifiable. Furthermore, the models of Table 3 suggest that the many decision heuristics and biases which arise from miscon-ceptions or misinterpretations could, in fact, be consequences of imperfect intra-personal communication. If so, the practical implication is that this class of decision error should arise consistently, for any individual in any affective state, unless that individual were to know of their susceptibility and consciously correct for it. Tversky & Kahneman (1974) provide a concise account of many heuristics and biases which might fall into this class of decision error, which future research in the spirit of Table 3 might therefore seek to explain.

The literature summarised in Tables 1 and 2 addresses the separate class of decision-making that can only be declared suboptimal in comparison to normative economic stan-

dards. That literature demonstrates that this second class of error could arise as the natural consequence of a conflict between deliberative and impulsive decision-criteria. Although the theoretical elegance of this explanation does not, in itself, validate it as even a stylised description of human cognition; authors such as Loewenstein & O'Donoghue (2005) and Brocas & Carrillo (2008) have derived a compelling case for its validity from the psychological and neurological literatures. The conflicting selves paradigm evidently also has intuitive appeal, since it forms the basis of the bestselling books of Kahneman (2011) and Peters (2012).

In providing a theoretical explanation for the biases of human decision-making, dual-self models can also explain those actions which might arise as a rational response to such errors. Indeed, the first dual-self model was created to explain the widespread use of apparently irrational savings commitment devices, and did so in essence by demonstrating that they constitute a sophisticated response to present-bias. The insight of Thaler & Shefrin (1981) was that an economically rational 'planner', who is aware that their actions in future periods will be controlled by a myopic 'doer', would optimally seek to constrain the expenditure of that future self. They envisaged simple rule-of-thumb constraints for pragmatic reasons, however it was later shown that an expenditure cap would also be the first-best solution for the 'planner' in a wide range of circumstances (Amador, Werning & Angeletos 2006). Brocas & Carrillo (2008) formally derived present-bias from this framework, and also demonstrated that the anomalously large correlation between current income and expenditure could itself represent an efficient rule-of-thumb. The main criticism of the approach taken by these Table 1 models is that they require a highly exogenous change in disposition between decisions. Although Gul & Pesendorfer (2001) elegantly resolve this inconsistency by axiomatising preferences over sets of alternatives, an elementary alternative resolution is provided by the approach of Table 4. If the decision-maker's state of mind is a probability distribution that is conditional upon the context of a decision, then the myopic 'doer' will be active with positive probability in present period consumption decisions, but indifferent in decisions between exclusively future alternatives.

It is remarkable that, although we cluster Tables 1 and 3 by their modelling approach, a taxonomy by implications would produce almost identical categorisations. This observation suggests that conflicting motivations and imperfect communication between selves might be important determinants of separate aspects of human behaviour. Only one existing paper challenges this conclusion – Brocas & Carrillo (2014b) derives decreasing impatience from dual selves with aligned preferences and asymmetric information. However they require some rather implausible assumptions to achieve this: both of their dual selves are deliberative, their 'computation' system is aware of the entire distribution of possible future consequences, and their 'modulation' system has foreknowledge of precisely which of those possible consequences will be realised. It would therefore be a

worthwhile objective for future research to investigate precisely which classes of decision errors can arise from the existing modelling approaches, under plausible assumption sets.

In contrast Tables 2a, and 2b are so-numbered because they produce comparable results through conceptually related modelling approaches. The results of Benhabib & Bisin (2005) and Brocas & Carrillo (2014b) are particularly similar, and when Loewenstein, O'Donoghue & Bhatia (2015) of 2b apply their model to temporal inconsistency it becomes a special case of the Fudenberg & Levine (2006) model of 2a. More significantly, it should be noted that the rational response to each situation modelled in Tables 1, 2a, and 2b is formally equivalent to the response generated by some single-self model. In particular, Loewenstein, O'Donoghue & Bhatia (2015) and Fudenberg & Levine (2006) derive their models' single-self equivalent within their respective papers, whilst Dekel, Lipman & Rustichini (2001) showed that all decisions generated by a model of type 1 can be reproduced by singular preferences over lotteries over sets of alternatives.

One could therefore claim that the models of Table Tables 1, 2a, and 2b merely represent an alternative interpretation of standard behavioural decision criteria. That claim is not strictly true. To illustrate, consider the costly willpower paradigm of Loewenstein & O'Donoghue (2005, 2015), where a rational agent must exert willpower to move away from their instinctive response towards a deliberative optimum. That framework can *derive* decision criteria such as the quasi-hyperbolic discounting of Laibson (1997), the altruist-egoist trade-off advocated by Becker (1976), or the kantian trade-off proposed by Alger & Weibull (2013), each of which are usually taken as primitive assumptions. Thus the costly willpower paradigm can provide an unified derivation for many behavioural descriptions of human decision-making. This observation is the chief motivation for the series of papers by Fudenberg and Levine (2006, 2011, 2012, 2014), which develop the costly willpower paradigm and apply it to explain several prominent decision anomalies.

The derivation of standard behavioural decision criteria through a costly willpower paradigm does, nevertheless, *induce* an alternative interpretation of the weighting parameter. For example, an agent's apparent overweighting of present consumption would become a consequence of their rational alleviation of the discomfort of resisting temptation. This interpretation offers distinct policy implications from the standard interpretation. For instance, Fudenberg & Levine (2012) model an agent with a dynamic stock of willpower, and Loewenstein & O'Donoghue (2005, p.42) discuss factors which could influence the present level of willpower to conclude that "the rich, who are not confronted with the constant task of reigning in their desires, are likely to judge the short-sighted behaviors of the poor too harshly." Such judgement may be justified under the standard economic paradigm, which suggests that poverty (at least in advanced economies) is the just deserts of impatient or lazy decision-making; however that simplistic view is challenged under the costly willpower paradigm, whereby individuals' willpower could be shaped during their childhood and depleted if constantly called upon to complete thank-

less and mundane tasks at work and at home. Embrey (2017) presents a dual-self model that develops this insight to explain the observed persistence of multiple dimensions of social inequality.

The models presented in Embrey (2017) and (Embrey & Kaivanto 2018) are unique in focussing on the dynamic consequences of decision-making under conflicting motivations, rather than calculating a rational agent's optimal response to that framework under the assumption of perfect self-knowledge. By allowing nature to determine which state of mind will predominate for any given decision-instance, these models also provides an economic description of decisions which the agent need not consciously recognise that they are taking. This heterogeneity in ways-of-thinking is possible since the state of mind distribution decision criteria is influenced both by an individual's human nature, and by the exogenous state of nature – an approach pioneered by Bernheim & Rangel (2004). This approach provides a descriptive theory of the imperfect and idiosyncratic decision-making which characterises Homo sapiens, and contrasts with the standard paradigm, wherein homogeneous agents optimally determine actions in response to some unified utility maximisation problem.

The models of Tables 1, 2a, and 2b also impose thought-process homogeneity; either by imposing an exogenous rule for the decision-maker's state of mind, or by assuming that agents control that first-stage decision through some meta-rational process. In contrast, the table 4 models of Laibson (2001) and Bernheim & Rangel (2004) explain many of the stylised facts of addiction through history-dependent utility functions, and through a stochastic state of mind which determines whether a craving will be triggered. That stochastic element is, in both cases, designed to capture the concept of cue-driven addiction, however it could also be interpreted more broadly to describe priming and framing effects. Interestingly, Bernheim & Rangel (2004, p.1572) stringently reject the dual-self interpretation of their model on the premise that addicts frequently consider their actions in an affective 'hot' state of craving to be a mistake. Thus, they claim, those actions cannot be the result of any underlying preferences. This claim contradicts the concept of revealed preference, and has little philosophical basis: Von Mises (1949) and even Bentham (1789) declared preferences to be situation-specific, and it is a key strength of dual-self theory that such specificity can be operationalised[7].

The models of Table 4 impose the fewest restrictions on the generalised decision framework, since circumstances, individuality, and chance are each allowed to influence the determination as to which decision criterion will predominate. This approach has two substantial benefits. first, it is the only approach to be compatible with a default-interventionist interaction between selves (Tables 2a, and 2b, and 3 require parallel processing which is rejected by the psychological and neuroscientific evidence – see, for example Evans & Stanovich (2013), and Bechara (2005)), and second, it provides economic theory with the opportunity to explore the implications of heterogeneity in ways-

of-thinking, and thereby also to describe actions which need not be the result of any concious decision. Such an approach is, however, under-exploited in the economic literature, wherein all individual differences are traditionally attributed to heterogeneity in tastes.

The synopsis in this section has suggested that diverse implementations of the dual-self paradigm can explain many human deviations from economically rational decision-making. Mathematically fallacious decision heuristics could arise from imperfect communication between selves, whilst normatively poor decision-making could arise from a conflict between deliberative and impulsive selves. These two classes of decision error are currently produced by the distinct modelling approaches of Tables 3 and 1 respectively, however it is a stated aim of many dual-self theorists to work towards an unified theoretical explanation of human behavioural anomalies (an aim eloquently advocated in Fudenberg 2006). The present review supports that aim by providing a convenient summary of the state of art in dual-self modelling, and by deriving a generalised decision framework which nests much of the dual-self literature.

Under that generalised framework, an agent's state of mind could be determined according to a probability distribution conditional upon circumstances, heterogeneity, and decision history. Little existing research maintains the full generality of that paradigm, however this may be a fruitful avenue for future research toward an unified theoretical understanding of behavioural decision-making. Consider, that if impulsive decision-making: overweights representativeness, updates imperfectly, and is cued by availability, then many of the heuristic class of error could be explained. If it also: overweights present desires, fears the unknown, and is cued by affective state, then many of the normatively suboptimal class of errors could be explained[8]. Since impulsive decision-making is, indeed, commonly attributed with qualities such as these (see, for example, Kahneman 2011), a generalised implementation of the quintessential human conflict between deliberative and impulsive reasoning might be shown to explain many of the behavioural anomalies in human decision-making.

# 4    Conclusion

This paper provides a holistic review of economic dual-self modelling. We find that the majority of the existing literature seeks to encapsulate the same human conflict between deliberative and impulsive reasoning, but that the approach taken to modelling the conflict between those decision criteria varies substantially. In order to taxonomise the dual-self literature, we therefore present a generalised decision framework in Section 3.1 that nests both Neoclassical and dual-self decision-theory. Existing approaches toward modelling the interaction between conflicting decision criteria include: a simple context-dependent rule, meta-rationality, game-theoretic interaction, and a conditional

probability distribution over the set of possible decision criteria. Tables 1-4 summarise the assumptions and implications of the economic dual-self literature, based upon that categorisation, and Section 3.2 reviews the main achievements and limitations of each modelling approach. We conclude that substantial new insights into normatively sub-optimal decision-making have been gained through the dual-self approach, and that the disparate strands of the existing literature could potentially be unified under the most general probabilistic dual-self approach.

Our generalised decision theory extends the standard Neoclassical model by including an explicit meta-level interaction between decision criteria. Superficially, one might there-fore suppose that dual-self theory is less parsimonious than the Neoclassical approach. In fact, the converse is true. Behavioural theories already include multiple utility compo-nents, but they commonly maintain the assumption that these can be 'spliced together' into a single decision criterion with neither justification nor acknowledgement. The meta-rational dual-self models summarised in Tables 2a and 2b require precisely the same as-sumption, but they make this explicit and base it upon psychological or neuroscientific insights. Moreover, the dual-self models of Table 1 and Table 4 require strictly weaker assumption sets. We provide a thorough comparison of the foundational assumptions of each approach in Section 2.

Ostensibly, these advances in dual-self decision-theory could constitute a challenge to the standard Expected Utility paradigm. This is not so. The generalised decision frame-work described in Section 3 is a strict generalisation of the Neoclassical paradigm, and moreover it is typically implemented by specifying two alternative utility maximisation problems. Furthermore, our discussion of the foundational assumptions and implications of the dual-self generalisation concluded that the Neoclassical approach is ideally suited to describe any situation which may be characterised by a single clear decision objective. It is only in 'behavioural' situations – those characterised by multiple, conflicting, moti-vations – that the dual-self generalisation provides greater insight. In such situations, an *as if* interpretation of expected utility theory can only describe the outcome of human decision-making, whereas the generalised approach proposes a descriptive theory of the decision-making process. Furthermore, the most generalised models admit heterogeneity in that process, which, given the wide-ranging applications of the literature reviewed here, may well prove to be an important determinant of individual differences.

Table 1: *Summary of dual-self models where the state of mind is determined by the type of decision under consideration*

| Papers | Decision-Making | Applied to | Implications |
|---|---|---|---|
| Thaler & Shefrin (1981) JPE | The 'planner' maximises lifetime utility by costlessly imposing rules in period 0; the 'doer' then makes myopic consumption decisions each period according to those rules. | Inter-temporal consumption; | Self-imposed rules-of-thumb likely; |
| Amador, Werning & Angeletos (2006) Econometrica | The 'principal' may restrict the set of alternatives which the 'agent' then chooses between. The latter has information on current tastes, but is also subject to temptation. That temptation may be overcome at a cost. | Financial decisions; Fiscal rules; Paternalism; Controlling externalities; | A minimum savings rule is optimal in a large class of situations; Restrictive laws such as minimum schooling and savings may be optimal; |
| Brocas & Carrillo (2008) AER | As above, but allows temptation both due to myopia and due to preference bias; also extends to interrelated decisions over multiple goods. | Inter-temporal consumption and labour decisions; Choice bracketing; | Time inconsistency; Self-imposed rules-of-thumb optimal; Narrow choice bracketing; Consumption over-responsive to current income. |

Table 2: *Summary of dual-self models where the state of mind is determined by expected-utility meta-rationality:*

Table 2a:  *Models where the deliberative self manipulates the preferences or information of the impulsive decision-maker at a cost*

| Papers | Decision-Making | Applied to | Implications |
|---|---|---|---|
| Fudenberg & Levine (2006) AER<br><br>Fudenberg & Levine (2011) AEJ: Micro<br><br>Fudenberg, Levine & Maniadis (2014) J. Econ. Psych. | The 'long-run self' maximises lifetime utility by paying a cost to manipulate the preferences of the 'short-run self', which makes myopic decisions. | Inter-temporal consumption; Procrastination; Cognitive load; | Time-inconsistency; Rabin's Paradox (inconsistent risk-aversion); Allais Paradox; Intransitivity; Cognitive load effects; Preference reversal for delayed payoffs; |
| Fudenberg & Levine (2012) Econometrica | As above, but weakens the myopia of the short-run self, and introduces a dynamically changing stock of cognitive resources. | Inter-temporal consumption; | Removes the discontinuities arising from arbitrary redefinition of periods; Smooths commitment decisions; |
| Brocas & Carrillo (2014*b*) Working Paper | The 'computation' system optimises consumption decisions given tastes and an estimate of the long-term consequences; the 'modulation' system monitors this and may provide information on the true long-term consequences at a cost. | Health decisions; Inter-temporal consumption; Addiction; | Decreasing impatience; Biased beliefs bias decision-making; Excessive desire biases decision-making. |
| Peysakhovich (2014) JEBO | As per (Fudenberg & Levine 2006). | Inter-temporal commitment devices; | 'Carrot' devices are preferred to 'sticks' and binding commitment, in the presence of either trembles or stochastic payoffs. |

*Table 2b:   Models where the deliberative self determines behaviour, by moving away from the impulsive optimum at a cost*

| Papers | Decision-Making | Applied to | Implications |
|---|---|---|---|
| Benhabib & Bisin (2005) Games and Econ. Behavior | 'Automatic' processing maximises gratification; 'controlled' processing overrides that decision iff it calculates that the cost of regret would outweigh a constant cost of exerting willpower. | Inter-temporal consumption; | Greater temptations increase both self-control and expenditure; Cognitive load effects; Simplistic rules-of-thumb may be optimal; 'controlled' decisions take longer; |
| Loewenstein, O'Donoghue & Bhatia (2015) Decision<br><br>Loewenstein & O'Donoghue (2005) Working Paper | The 'affective' system maximises gratification; The 'deliberative' system maximises 'rational' utility less a willpower cost of moving away from the affective optimum. | Inter-temporal choice; Risk attitudes; Social preferences; | Environmental, framing, and priming effects; Effect of cognitive load; Possible decision-making mechanism; Time-inconsistency; Temptation cost. |

Table 3: *Summary of dual-self models where the state of mind is determined by game-theoretic interaction*

| Papers | Decision-Making | Applied to | Implications |
|---|---|---|---|
| Brocas & Carrillo (2012) Games and Econ. Behavior | A 'sensory' function imperfectly encodes the state of nature; a 'decision' system then selects actions in a coordination game with nature by optimally partitioning the image of the sensory function. | Information processing; Belief formation; Bounded cognition; | Belief anchoring; Belief polarisation; Belief divergence, since preferences shape beliefs; Unlikely alternatives are disregarded; |
| Cunningham (2013) Working Paper | 'System 1' and 'System 2' have identical preferences but each perceives private information. System 1 forms a judgement, then System 2 decides all actions taking that judgement into account. | Decision heuristics and biases; | Projection bias; Framing effects; Judgements may appear inconsistent, but biases will be systematic; Inconsistencies will vanish if problems are solved jointly; Biases will reduce with experience; |
| Alonso, Brocas & Carrillo (2014) Rev. Econ. Stu. | Multiple 'agents' require privately known levels of cognitive resources to perform tasks; a single 'principle' allocates limited resources to optimise global performance weighted by task importance. | Cognitive processing; | Performance deteriorates when task complexity or cognitive load exceed some threshold; Inertia in ways of thinking; Optimal task allocation reminiscent of microprocessor threading. |

Table 4: *Summary of dual-self models with a probabilistic State-of-Mind, conditional upon: circumstances, heterogeneity, decision-history, ...*

| Papers | Decision-Making | Applied to | Implications |
|---|---|---|---|
| Laibson (2001) QJE (not framed as a dual-self theory) | Nature determines which of two states will apply. Each state is characterised by rational action, based upon payoffs which are affected by the history of decisions made in that state, such that habit-formation takes place. | Addiction; Habit formation; Marketing; | Commitment and cue-management are optimal; Good-specific time-inconsistency; Stochastic binging; Polarisation of substance use; Marketing could cue desire; |
| Bernheim & Rangel (2004) AER (reject the dual-self interpretation, so that 'cold' preferences can be interpreted as welfare optimising, and 'hot' decisions as systematic errors.) | Each period the 'cold' self rationally chooses a lifestyle activity, which, together with their addiction history, affects the likelihood of cueing their 'hot' self. The hot self then consumes substances with certainty, whereas the cold self consumes iff it is optimal to do so. | Addiction; | Polarisation of use; Welfare loss iff addicts want to (but fail to) quit; Hence addiction not due to poor information; Commitment and cue-management could break addiction; Marketing could cue desire; |
| Embrey (2017) Working Paper | The 'deliberative' self will overrule the 'impulsive' self with an individual- and situation-specific probability which depends upon agents' characteristics, background, decision history, and the present circumstances. | Human capital development; Life-course decisions; | Objectively poor decision-making; Non-cognitive ability ≈ one's propensity to act deliberatively; Divergent life courses due to early-life circumstances and experiences; Underlying mechanisms suggest possible interventions; |
| Embrey & Kaivanto (2018) Working Paper | Decisions will be made according to: 'deliberative', 'behavioural', 'impulsive', or 'automatic' choice criteria; similar mechanism to above. | Social engineering attacks (e.g. phishing); | Inconsistency between training and true susceptibility; Stepping-stone attacks; Training should focus on meta-cognition. |

# Acknowledgements

# Notes

1. For an exposition of the dual-process paradigm see Kahneman & Frederick (2002); for a review of psychological theories based upon it see Alós-Ferrer & Strack (2014); for a discussion of its neurological and evolutionary justifications see Cohen (2005).

2. Bentham (1789) is also frequently attributed with the postulates that individuals' utility valuations ought to be quantitatively comparable, and that individuals' experienced utility is identically their decision utility. Although both of these postulates are problematic, neither is relevant to our present discussion (which concerns individual agents' decision utilities), except that the model proposed here does provide a candidate explanation for the empirical discrepancy between individuals' predicted and experienced utilities (see Kahneman, Wakker & Sarin 1997). It should also be noted that the latter attribution is false.

3. Homo sapiens do, however, have a tendency to supply ex post facto rationalisations for our behaviour (Loewenstein 1996). That tendency, known to psychologists as hindsight bias, may contribute to the dominance of the 'single-self' paradigm in microeconomic theory.

4. For example: Loewenstein & O'Donoghue (2005) contrast deliberative with affective motivations, Strack & Deutsch (2004) contrast reflective with impulsive motivations, Bernheim & Rangel (2004) contrast cold with hot decision criteria, and Thaler & Shefrin (1981) describe a planner and a doer. Though psychology and neuroscience rightly scrutinise the subtle distinctions between these concepts (See, for example, Evans 2008), we shall concern ourselves with the substance rather than the nomenclature of the dual-self literature.

5. For a full theoretical analysis of such decisions see Embrey (2017).

6. Multiple selves are considered by Alonso, Brocas & Carrillo (2014) and by Embrey & Kaivanto (2018).

7. Nothing in Bernheim and Rangel's otherwise compelling paper relies on this claim; it is merely a convenience which removes the need to justify the use of 'cold' preferences as a welfare criterion.

8. It is interesting to observe that xenophobia – accurately 'fear of the unknown' – could underpin not only risk aversion, but also present-bias, habit-formation, and discrimination.

# References

**Alger, Ingela, and Jörgen W. Weibull.** 2013. "Homo Moralis — Preference Evolution Under Incomplete Information and Assortative Matching." *Econometrica*, 81(6): 2269–2302.

**Alonso, Ricardo, Isabelle Brocas, and Juan D. Carrillo.** 2014. "Resource allocation in the brain." *Review of Economic Studies*, 81(2): 501–534.

**Alós-Ferrer, Carlos, and Fritz Strack.** 2014. "From dual processes to multiple selves: Implications for economic behavior." *Journal of Economic Psychology*, 41: 1–11.

**Amador, Manuel, Iván Werning, and George-Marios Angeletos.** 2006. "Commitment vs . Flexibility." *Econometrica*, 74(2): 365–396.

**Baumeister, Roy F.** 2003. "The Psychology of Irratioonality: Why People Make Foolish, Self-defeating Choices." In *The Psychology of Economic Decisions: Rationality and Well-being*Vol. 1, ed. Isabelle Brocas and Juan D Carrillo, Chapter 1, 3–16. Oxford: Oxford University Press.

**Bechara, Antoine.** 2005. "Decision making, impulse control and loss of willpower to resist drugs: a neurocognitive perspective." *Nature: Neuroscience*, 8(11): 1458–1463.

**Becker, Gary S.** 1976. "Altruism, egoism, and genetic fitness: Economics and sociobiology." *Journal of Economic Literature*, 14(3): 817–826.

**Benhabib, Jess, and Alberto Bisin.** 2005. "Modeling internal commitment mechanisms and self-control: A neuroeconomics approach to consumption-saving decisions." *Games and Economic Behavior*, 52(2): 460–492.

**Bentham, Jeremy.** 1789. "An Introduction to the Principles of Morals and Legislation." In *The Collected Works of Jeremy Bentham, 1970*, ed. J.H. Burns and H.L.A. Hart. London: The Althlone Press.

**Bernheim, Douglas B, and Antonio Rangel.** 2004. "Addiction and Cue-Triggered Decision Processes." *The American Economic Review*, 94(5): 1558–1590.

**Brocas, By Isabelle, and Juan D Carrillo.** 2008. "The Brain as a Hierarchical Organization." *The American Economic Review*, 98(4): 1312–1346.

**Brocas, Isabelle.** 2012. "Information processing and decision-making: Evidence from the brain sciences and implications for economics." *Journal of Economic Behavior & Organization*, 83(3): 292–310.

**Brocas, Isabelle, and Juan D Carrillo.** 2003. *the Psychology of Economic Decisions: Rationality and Well-being.* Vol. 1 Oxford: Oxford University Press.

**Brocas, Isabelle, and Juan D. Carrillo.** 2012. "From perception to action: An economic model of brain processes." *Games and Economic Behavior*, 75(1): 81–103.

**Brocas, Isabelle, and Juan D. Carrillo.** 2014a. "Dual-process theories of decision-making: A selective survey." *Journal of Economic Psychology*, 41: 45–54.

**Brocas, Isabelle, and Juan D. Carrillo.** 2014b. "Value computation and value modulation : a dual-process theory of self-control." *Unaffiliated working paper*. Available at: http://www.neuroeconomictheory.org/PDF/temptation.pdf.

**Cohen, Jonathan D.** 2005. "The vulcanization of the human brain: a neural perspective on interactions between cognition and emotion." *Journal of Economic Perspectives*, 19(4): 3–24.

**Crawford, Vincent P, Miguel A Costa-Gomes, and Nagore Iriberri.** 2013. "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications." *Journal of Economic Literature*, 51(1): 5–62.

**Cunningham, Tom.** 2013. "Biases and Implicit Knowledge." *Unaffiliated working paper*. Available at: https://www.parisschoolofeconomics.eu/IMG/pdf/cunningham_paper.pdf.

**Dekel, B Y Eddie, Barton L Lipman, and Aldo Rustichini.** 2001. "Representing Preferences with a Unique Subjective State." *Econometrica*, 69(4): 891–934.

**Donohew, Lewis, Rick Zimmerman, Pamela S Cupp, Scott Novak, Susan Colon, and Ritta Abell.** 2000. "Sensation seeking, impulsive decision-making, and risky sex: implications for risk-taking and design of interventions." *Personality and Individual Differences*, 28(6): 1079–1091.

**Edwards, Ward.** 1954. "The theory of decision making." *Psychological Bulletin*,

51(4): 380–417.

**Ellsberg, Daniel.** 1961. "Risk, Ambiguity, and the Savage Axioms." *The Quarterly Journal of Economics*, 75(4): 643–669.

**Embrey, Iain Peter.** 2017. "States of Nature and States of Mind: A Generalised Theory of Decision-Making, evaluated by application to Human Capital Development." *Lancaster University Economics Working Paper, No.2017/032.*

**Embrey, Iain Peter, and Kim Kaivanto.** 2018. "Many Phish in the C: A Coexisting-Choice-Criteria Model of Security Behavior." *Lancaster University Economics Working Paper, No. 2018/024.*

**Evans, Jonathan St B T.** 2008. "Dual-processing accounts of reasoning, judgment, and social cognition." *Annual review of psychology*, 59: 255–278.

**Evans, Jonathan St. B. T., and Keith E. Stanovich.** 2013. "Dual-Process Theories of Higher Cognition." *Perspectives on Psychological Science*, 8(3): 223–241.

**Fudenberg, By Drew, and David K Levine.** 2011. "Risk , Delay , and Convex Self-Control Costs." *American Economic Journal: Microeconomics*, 3(3): 34–68.

**Fudenberg, Drew.** 2006. "Advancing beyond "Advances in Behavioral Economics"." *Journal of Economic Literature*, 44(3): 694–711.

**Fudenberg, Drew, and David K. Levine.** 2006. "A Dual-Self Model of Impulse Control." *The American Economic Review*, 96(5): 1449–1476.

**Fudenberg, Drew, and David K Levine.** 2012. "Timing and Self-Control." *Econometrica*, 80(1): 1–42.

**Fudenberg, Drew, David K. Levine, and Zacharias Maniadis.** 2014. "An approximate dual-self model and paradoxes of choice under risk." *Journal of Economic Psychology*, 41: 55–67.

**Gilboa, I., and D. Schmeidler.** 1995. "Case-Based Decision Theory." *The Quarterly Journal of Economics*, 110(3): 605–639.

**Goldman, David.** 2012. *Our Genes, Our Choices.* London: Elsevier.

**Graham, Hilary.** 2007. *Unequal Lives: Health and Socioeconomic Inequalities.* Maidenhead: McGraw-Hill Education.

**Gul, Faruk, and Wolfgang Pesendorfer.** 2001. "Temptation and Self-Control." *Econometrica*, 69(6): 1403–1435.

**Kahneman, Daniel.** 2011. *Thinking, Fast and Slow.* London: Penguin Books Ltd.

**Kahneman, Daniel, and Shane Frederick.** 2002. "Representativeness revisited: Attribute substitution in intuitive judgment." In *Heuristics and Biases: The Psychology of Intuitive Judgement*, ed. T. Gilovich, D. Griffin and D. Kahneman, Chapter 2, 49–81. Cambridge: Cambridge University Press.

**Kahneman, Daniel, Peter P. Wakker, and Rakesh Sarin.** 1997. "Back to Bentham? Explorations of Experienced Utility." *Quarterly Journal of Economics*, 112(2): 375–405.

**Kosfeld, Michael, Ernst Fehr, Jorgen Weibull, and Topi Miettinen.** 2017. "Revealed Preferences in a Sequential Prisoners' Dilemma: A Horserace Between Five Utility Functions." *Unaffiliated working paper.* Available at: https://sites.google.com/site/joergenweibull/working-papers.

**Laibson, David.** 1997. "Golden Eggs and Hyperbolic Discounting." *The Quarterly Journal of Economics*, 112(2): 443–477.

**Laibson, David.** 2001. "A Cue-Theory of Consumption." *The Quarterly Journal of Economics*, 116(1): 81–119.

**Loewenstein, George.** 1996. "Out of control: Visceral influences on behavior." *Orga-*

*nizational Behavior and Human Decision Processes*, 65(3): 272–292.

**Loewenstein, George, and Ted O'Donoghue.** 2005. "Animal spirits: Affective and deliberative processes in economic behavior." *Unaffiliated working paper.* Available at: https://blogs.cornell.edu/odonoghue/files/2015/10/will5_05-227fjlg.pdf.

**Loewenstein, George, Ted O'Donoghue, and Sudeep Bhatia.** 2015. "Modeling the interplay between affect and deliberation." *Decision*, 2(2): 55–81.

**Mani, Anandi, Sendhil Mullainathan, Eldar Shafir, and Jiaying Zhao.** 2013. "Poverty Impedes Cognitive Function." *Science*, 341(6149): 976–980.

**Marshall, Alfred.** 1890. *Principles of Economics.* London: MacMillan and Co. Ltd.

**Peters, Steve.** 2012. *The Chimp Paradox: The mind Management Programme for Confidence, Success, and Happiness.* London: Ebury Publishing.

**Peysakhovich, Alexander.** 2014. "How to commit (if you must): Commitment contracts and the dual-self model." *Journal of Economic Behavior and Organization*, 101: 100–112.

**Rabin, Matthew.** 2000. "Risk Aversion and Expected-utility Theory: A Calibration Theorem." *Econometrica*, 68(5): 1281–1292.

**Rabin, Matthew.** 2002. "A perspective on psychology and economics." *European Economic Review*, 46(4-5): 657–685.

**Raiffa, Howard.** 1969. "Preferences for Multi-Attributed Alternatives." *Memorandum, No. RM-5868-DOT/RC.* Santa Monica: The RAND Corporation.

**Samuelson, Paul A.** 1938. "A Note on the Pure Theory of Consumers' Behaviour." *Economica*, 5(17): 61–71.

**Savage, Leonard J.** 1954. *The Foundations of Statistics.* New York: John Wiley & Sons Inc.

**Schelling, Thomas C.** 1984. "Self-Command in Practice, in Policy, and in a Theory of Rational Choice." *The American Economic Review*, 74(2): 1–11.

**Smith, Adam.** 1759. *The Theory of Moral Sentiments.* London: A. Millar.

**Strack, Fritz, and Roland Deutsch.** 2004. "Reflective and impulsive determinants of social behavior." *Personality and social psychology review*, 8(3): 220–247.

**Thaler, Richard H., and H.M. Shefrin.** 1981. "An Economic Theory of Self-Control." *Journal of Political Economy*, 89(2): 392–406.

**Tversky, A., and D. Kahneman.** 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science*, 185(4157): 1124–1131.

**Von Mises, Ludwig.** 1949. *Human Action: A Treatise on Economics.* London: William Hodge and Company Ltd.

**von Neumann, Ludwig Johann, and Oskar Morgenstern.** 1947. *Theory of Games and Economic Behavior.* 2 ed. Princeton: Princeton University Press.

**Watt, Richard Geddie.** 2007. "From victim blaming to upstream action: Tackling the social determinants of oral health inequalities." *Community Dentistry and Oral Epidemiology*, 35(1): 1–11.