***In Defence of Individual Rationality[[1]](#footnote-1)\****

**Abstract**: Common-sense (or folk) psychology holds that (generally) we do what we do for the reasons we have. This common-sense approach is embodied in claims like “I went to the kitchen because I wanted a drink” or “She took a coat because she thought it might rain and hoped to stay dry”. However, the veracity of these common-sense psychological explanations has been challenged by experimental evidence (primarily from behavioural economics and social psychology) which appears to show that individuals are systematically irrational – that often we do not do what we do because of the reasons we have. Recently, some of the same experimental evidence has also been used to level a somewhat different challenge at the common-sense view, arguing that the overarching aim of reasoning is not to deliver better or more logical decisions for individual reasoners, but to improve group decision making or to protect an individual’s sense of self. This paper explores the range of challenges that experimental work has been taken to raise for the common-sense approach and suggests some potential responses. Overall, I argue that the experimental evidence should not (currently) lead us to a rejection of individual rationality.

**Current word count (exc. abstract): 7566**

According to a fairly standard view in philosophy, intentional actions are the result of rational processing of reasons: I reach for the jar because I *want a cookie* and *believe there is a cookie in the jar*.[[2]](#footnote-2) But how widely applicable is this sort of belief-desire understanding of action? Recent work in social psychology and behavioural economics suggests the answer is ‘not very widely’, for experimental work apparently shows that very often what we do is *not* a rational reflection of our beliefs and desires. Instead, we act out of intuitive gut-feelings, using cognitive short-cuts or heuristics which result in us being blown this way and that by situational features that do not constitute rational causes for action. The common-sense view of individuals, as generally acting in light of their reasons, turns out not to survive confrontation with the facts about how we actually decide what to do. In this paper I explore the experimental evidence offered for this claim and suggest a defence of individual rationality is in order.

**1. Two arguments against individual rationality**

To appreciate the arguments against individual rationality, we first need to get a somewhat clearer view of the position these arguments oppose: what is it for an individual to count as making a rational decision in the first place? While there is much to be said in answer to this question, for the purposes of this paper I’ll simply set out the view that opponents of individual rationality seem to accept, which we might call ‘Classical Rational Choice Theory’. [[3]](#footnote-3) According to this approach, to count as rational individual decision-making needs to:

* aim at maximising utility
* consider all the evidence
* utilise classical logic in deterministic environments and probability theory in uncertain/vague situations.

For instance, capturing the first two points, Ariely 2008: xxx writes that “the assumption that we are rational implies that, in everyday life, we compute the value of all possible options we face and then follow the best possible path of action”. This Classical model has been very influential, underpinning the view of humans as *homo economicus* in behavioural economics and *homo politicus* in political theory, where agents are viewed as primarily self-interested individuals capable of weighing up all the available evidence in order to arrive at an ideal solution to any problem or decision they face. I’ll return to the question of whether opponents are right to take Classical Rational Choice Theory to characterise human rationality below, but, given the view, the challenge is clear. For there is a wealth of experimental evidence to show that people’s actual decisions in a range of cases don’t in fact line up with the predictions of Classical Rational Choice Theory.[[4]](#footnote-4) Individual decision making thus turns out to be frequently non-rational.[[5]](#footnote-5)

 It is important to notice, however, that, on the basis of (often the same) experimental evidence, opponents offer two rather different arguments for this conclusion:

1. Use of automatic decision-making systems: agents often *don’t try to reason* at all (instead relying on non-logical, automatic decision-making processes).
2. Improper use of a logical system: agents often *try but fail* *to reason properly* (although utilising deliberative processes, agents don’t meet the demands of rationality due to some systemic failure).

I’ll consider each of these challenges to individual rationality in turn, sampling the evidence offered for each and suggesting some reasons to be sceptical about the conclusions opponents have reached.[[6]](#footnote-6)

**2. Automatic (intuitive) decision-making systems**

As is well-known, many theorists have proposed that human decision-making is underpinned by two radically different systems or processes.[[7]](#footnote-7) On the one hand we have a slow, reflective, logical decision-making system – a broadly rational, belief-desire based system of the kind familiar to philosophers from discussions about ‘folk’ or ‘common-sense’ psychology. The outcomes of these processes are actions performed on the basis of reasons. On the other hand, however, we also have a fast, automatic, non-logical decision-making process – a system of intuition and habit which operates on the basis of quick and dirty ‘heuristics’, sacrificing accuracy for speed – which Kahneman 2011 speaks of as “a machine for jumping to conclusions”.

 To clarify the notion of a cognitive heuristic, Tversky & Kahneman’s seminal 1974 *Science* paper opens with discussion of a visual heuristic: we assess object distance in terms of clarity, so that we tend to think things we can see clearly are close, while those which are blurry are further away.[[8]](#footnote-8) In general, this method for assessing object distance works tolerably well – it doesn’t yield precise or fine-grained knowledge of distance, but it tends to permit a broadly accurate ranking of objects over distance. However, in conditions where clarity is either particularly good or particularly bad, the heuristic leads to errors – e.g. we tend to think that objects seen in poor light or rain are further away than they actually are because we equate being blurry with being further away, forgetting to adjust for lighting conditions. This example highlights three key features of heuristic thinking. First, heuristic thinking is thinking which neglects some information or options, looking only at a subset of what is relevant. Second, and connected to this, heuristic judgements are intuitive (they seem to be the result of quick, automatic triggering rather than slow, logical deliberation). Finally, and again related to the previous two properties, and as the visual case makes clear, the use of a heuristic, while practically useful, can lead us astray (e.g. by leading us to misrepresent object distance in certain lighting conditions). Since the Automatic System uses error-inducing heuristics instead of operating in line with Classical Rational Choice Theory, actions caused via its operation do not constitute rational actions.

 The suggestion (at least amongst advocates of the ‘default-interventionist’ approach, see Evans & Stanovich 2013) is that quick and dirty automatic processes (System 1) provide our default response to decision-making problems, with slow, logical processes (System 2) only being wheeled in when someone becomes aware that their automatic processes have made a mistake. So, why should we accept this ‘Dual Process’ model, contrasting as it does logical and automatic decisions? The evidence comes from two main sources: the ‘Cognitive Reflection Test’ and work from Kahneman and Tversky (amongst many others) on the range of cognitive heuristics and biases which apparently infect our decision-making.

***2.1 The Cognitive Reflection Test (CRT)***

In the Cognitive Reflection Test (Frederick 2005) subjects are asked to make a decision or judgement on an apparently simple question. In each case there is an answer that comes quickly, easily and naturally to mind. However, this intuitive answer can subsequently either be shown to be, or comes to be appreciated by the agent as, wrong.[[9]](#footnote-9) The original three cases which Frederick presented are now extremely famous:

1. A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball.
How much does the ball cost?
\_\_\_\_\_ cents
2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?
\_\_\_\_\_ minutes
3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?
\_\_\_\_\_ days

The intuitive answers to these questions are: 10 cents, 100 minutes, and 24 days. However, the correct (or logical) solutions are: 5 cents, 5 minutes, and 47 days. What Frederick found was that the majority of participants give the intuitive answers and not the logical answers to these questions, apparently demonstrating our Automatic System in action.[[10]](#footnote-10)

***2.2 Kahneman & Tversky’s Heuristics and Biases***

Tversky and Kahneman 1974 introduce their project by focusing on three examples of cognitive heuristics which people commonly deploy in decision-making, but which lead subjects into error in certain settings. These three original heuristics are:

1. Representativeness: assessing the likelihood of x being F via x’s similarity to other F’s.
2. Availability: assessing the likelihood of F occurring via the subject’s familiarity with other instances of F.
3. Anchoring & adjustment: assessing the likelihood or value of x from an accepted contextual anchor point, with adjustments made from that point.

In subsequent work, Kahneman and Tversky, and others working in the heuristics and biases programme, went on to add many further heuristics to the list.[[11]](#footnote-11) While surveying every suggested heuristic and bias would take us too far afield, two further influential proposals are worth noting:

1. Framing effects: “the large changes of preferences that are sometimes caused by inconsequential variations in the wording of a choice problem” (Kahneman 2011: 272)
2. Overconfidence: “our excessive confidence in what we believe we know, and our apparent inability to acknowledge the full extent of our ignorance and the uncertainty of the world we live in” (Kahneman 2011: 13-4)

 To see heuristics in action Kahneman and Tversky designed experimental probes where logical answers and intuitive answers come apart. As with the CRT, what they found was that a majority deliver the intuitive answer, demonstrating the operations of the Automatic System. So, for instance, the use of Representativeness was demonstrated via the (now famous) vignette:

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which is more probable?

1. Linda is a bank teller.
2. Linda is a bank teller and is active in the feminist movement.

What Tversky and Kahneman found was that a majority of participants answer (ii). Yet clearly this cannot be the logical answer, for a conjunction of claims ‘A&B’ can never be *more probable* than one conjunct, ‘A’, alone. As Thaler and Sunstein 2008: 29 write:

The error stems from the use of the representativeness heuristic: Linda’s description seems to match ‘bank teller and active in the feminist movement’ far better than ‘bank teller’. As Stephen Jay Gould…once observed, ‘I know [the right answer], yet a little homunculus in my head continues to jump up and down, shouting at me – “But she can’t just be a bank teller; read the description!”’ Gould’s homunculus is the Automatic System in action.

Although space prevents full presentation of the experiments run in support of the other heuristics, the general model is the same in each case. For instance, to demonstrate Availability in operation participants can be shown to overstate the likelihood of a given rare event (like a terrorist attack) when a past attack is made particularly salient for them (e.g. via newspaper coverage). Or again, in a Framing case, if participants are asked whether they would opt to have a hypothetical medical procedure, described in one scenario in terms of having ‘an 80% chance of survival’ and in another scenario as having a ‘20% chance of death’, they reliably opt for the procedure more in the first protocol, described in terms of survival.[[12]](#footnote-12) Yet an 80% chance of survival seems equivalent to a 20% chance of death, so the difference in response rates (due solely to the different framing of the same information) seems irrational. For each heuristic, then, experimental evidence supports the idea that we tend to deliver quick, intuitive responses to problems and that these intuitive responses often fail to match the judgements demanded by Classical Rational Choice Theory. Rather than reason our way to logical decisions, people deploy fast but inaccurate heuristics which work okay in general but which lead us into predictable errors in a variety of situations (Ariely 2008).

 So, should we accept this conclusion, jettisoning the idea that individuals generally make decisions in line with the demands of rationality, instead relying on intuitive gut feels? I think there are reasons to be cautious. First, it’s unclear that the experimental evidence in question really supports scepticism about rational decision-making. Second, I think there are reasons to reject the overarching Dual Process framework to which this challenge belongs.

**2.3 Challenging the experimental evidence for the Automatic System**

Advocates of individual rationality might seek to challenge the Dual Process framework by rejecting the experimental evidence used to support it. For a start, as with much work in social psychology, questions can be asked about how well some of the (often most surprising or interesting) experimental results replicate.[[13]](#footnote-13) However pointing to the Replication Issue doesn’t provide a full response, since many of the key results (such as those discussed above) do, in fact, replicate robustly. An alternative avenue for sceptics about the evidence to explore, then, involves concerns about ‘ecological validity’ – i.e. how closely experimental settings mirror real-world scenarios. For it turns out that improving the ecological validity of experiments (modelling situations more closely on real-life decision-making rather than abstract lab-based protocols) often raises the rate at which participants give logical (over intuitive) answers.[[14]](#footnote-14) However, the challenge I want to focus on here *accepts* (at least some of) the experimental findings at face value but suggests that these results can be accommodated by advocates of individual rationality.[[15]](#footnote-15) In particular, I want to suggest that a better sensitivity to the pragmatics of experimental prompts reveals that in many cases participants are in fact behaving far more logically than opponents give them credit for.

 So, take the classic CRT ‘bat and ball’ case (which I take to be the most compelling of the CRT questions). Here 5c is the correct answer *only if* the experimental prompt is heard as asking a question about the *relative value* of the two items. Yet this is a very unusual kind of query. A much more standard way, I suggest, to hear ‘The bat costs $1 more than the ball’ is as telling you what you have to pay to acquire the bat *after buying the ball*. But if the experimental prompt is heard in that way, then the correct answer to ‘How much is the ball?’ is in fact ‘10c’, just as most people say. I contend that what is going on here is not, as advocates of the CRT suggest, mass irrationality as a result of the operations of the Automatic System, but rather the simple fact that a majority of people hear the experimental prompt in the more practical way just suggested (while a minority of people are sensitive to the ‘relative value’ reading and thus give the ‘5c’ answer).

 To reinforce the idea that a better understanding of pragmatics can accommodate the evidence without any appeal to irrationality consider the infamous case of Linda. Here the fact that the speaker bothers to spend time telling participants that Linda fits the stereotype for a feminist is itself informative and it affects what the speaker is held to convey (Hertwig & Gigerenzer 1999, Mosconi & Macchi 2001). Experimental participants, no less than those engaged in ordinary conversations, are trying to grasp what the speaker is saying to them and in the Representativeness vignette it is clear that a reasonable conclusion is that the speaker intends to imply that Linda is a feminist (after all, why would they bother telling you all the stuff about her activism if not?). Yet if that is how the experimental materials are heard, then answering (ii) is not irrational, rather it displays a highly reasonable sensitivity to the pragmatics of conversation. Finally, consider our example of framing effects, where participants were more likely to choose a procedure described as having an ‘80% chance of survival’ over ‘20% chance of death’.[[16]](#footnote-16) First, note that there is a debate to be had about whether these two descriptions are in fact *semantically* equivalent (something required by any claim of irrationality based on non-equivalent responses), but, second, even waiving that concern, it is clear that the two claims are far from *pragmatically* equivalent. Framing things in terms of the positive ‘80% survival’ pragmatically conveys a hopeful message (i.e. that 80% is a good chance of survival in the circumstances), while choosing to focus on the negative (chance of death) pragmatically conveys that the speaker takes the risks to be particularly prominent. A number of theorists working on framing have developed precise mechanisms to explain how these pragmatic readings come about (for discussion, see Fisher & Mandel 2021) and I don’t want to take a stand on the precise model to be preferred, but the general claim that framing effects arise due to pragmatic differences in content deserves, I think, serious consideration. In general, then, I suggest that a better appreciation of the fact that experimental participants, just like other conversational participants, respond to what is pragmatically expressed by the speaker not just what is semantically conveyed, can help deflate allegations of irrationality based on experimental evidence.

**2.4 Challenging the existence of the Automatic System**

A second worry with arguments based on the operations of the Automatic System turns on whether the system really exists (in contrast to a Logical System) at all. For it seems, on closer inspection, that there is no property this system could have that is capable of *both* holding apart automatic and logical thinking *and* underpinning a claim of systematic or predictable irrationality.[[17]](#footnote-17) To see this, consider the range of properties appealed to above (and in the literature) when introducing the Automatic System, where its operations are held not to be rule based, to be fast, unconscious, effortless or automatic.[[18]](#footnote-18) Yet it seems none of these properties are capable of differentiating two different systems.

 So, for instance, it seems clear that Automatic thinking cannot be characterised as thinking which is not rule-based. For, first, the Automatic System is generally held to be computationally tractable, which requires it to be rule-based. Second, the examples of heuristics we have looked at pretty clearly constitute rules. For instance, the visual heuristic ‘judge objects which appear blurry to be far away’, and the Representativeness heuristic ‘assess the likelihood of x being F via x’s similarity to other F’s’, are both perfectly good rules, even if they are approximating rather than exact. So, the presence or absence of rules can’t be what differentiates the two kinds of system (a point acknowledged by Evans and Stanovich 2013: 231 who write “Evidence that intuition and deliberation are both rule-based cannot, by any logic, provide a bearing one way or the other on whether they arise from distinct cognitive mechanisms”). Alternatively, then, take the idea that Automatic thinking is fast, while logical thinking is slow (an idea which has shaped the entire debate in this area, as evidenced by the title of Kahneman’s bestselling *Thinking Fast and Slow)*. Yet it simply seems mistaken to hold that heuristic thinking is always fast or that logical thinking is always slow. Asked the question ‘15+5 = ?’, or given a really simple logic puzzle (like ‘If Maya had ice cream then she didn’t get sweets. Maya had ice cream. Did she get sweets?’), participants are likely (assuming they think the questions are asked seriously) to be able to deliver a logical answer extremely quickly.[[19]](#footnote-19) Furthermore, the idea that logical answers can sometimes be as fast as non-logical ones is in fact already embedded in the experimental evidence to hand, for even in CRT cases experimenters found that a minority of people give logically correct answers as their immediate, fast responses. The same goes for the claim that non-logical responses are automatic or effortless, while logical answers are, in contrast, controlled or effortful. Although it is likely that the phenomenology of decision-making differs across decision-makers, across problems, and across contexts, I suspect that few feel the need to exert more control to arrive at the answer ‘20’ when presented with ‘15+5 =?’ than when giving the answer that ‘Linda is feminist and a bank teller’ in the classic Representativeness experiment. Instead, I suggest, all these kinds of properties (speed, automaticity, conscious control) are ones that cross-cut any differences in the actual mechanisms of decision-making, with both heuristic based decisions and logical decisions sometimes being fast and sometimes being slow, etc.

 Finally, then, perhaps the advocate of a Dual Process model could simply appeal directly to the idea that while some decisions are based on the rules of logic others are based on different, non-logical transitions. In this way the Automatic System would be defined simply through its opposition to the Logical System (Sloman 1996). However, while this seems eminently plausible (since it seems highly unlikely that *every* decision we reach runs by way of classical logic or probability theory) it is a further step to the claim that this variability in methods entails irrationality. This takes us back to the assumption of §1 that Classical Rational Choice Theory provides the correct model of human rationality. For although the use of heuristics might not count as logical (though see Kruschke 2008 for an argument that even associative thinking should be treated as a form of probabilistic reasoning) this on its own does not show such thinking is particularly problematic, i.e. that it leads us into regular error. As Gigerenzer and colleagues have argued at length, the use of heuristics often yields *better* results than the application of logical rules to decision-making problems. Yet, if this is right, then it seems perverse to count use of a heuristic as irrational simply because it doesn’t fit within the constraints of Classical Rational Choice Theory.

 Neither logical nor heuristic rules *guarantee* that we get things right but a guarantee of infallibility would clearly be too high a demand to place on rationality. And if infallibility isn’t the bar, then it seems logic and heuristic thinking sit on all fours together, both potentially yielding the kind of *good enough* solutions we ought to demand of rational thinking. What this points to is the need (anticipated in §1) to refine Classical Rational Choice theory in order to make it suitable as an account of human thinking. We should allow that individuals are rational if they:

1. Aim at *good enough* solutions (‘satisfycing’) rather than maximizing utility (a point recognised long ago in Simon’s 1957 notion of ‘bounded rationality’).
2. Look at some but not necessarily all the evidence. (Gigerenzer & Gaissmaier 2011).
3. Use a reasoning process appropriate to the problem (drawing from classical logic, probability theory *and* associative thinking).

With these kinds of relaxations on the standards for rationality in hand, together with a recognition of the role of pragmatic understanding in experimental contexts, I think we can reject the claim that individuals typically fall short of rational decision-making. [[20]](#footnote-20)

 However, at this juncture a second line of attack hooves into view, according to which (contra the objection of §2) we do *try* to reason rationally but our attempts are prone to systematic failures. And these failures entail, just as advocates of the Automatic System suggest, that individuals often fail to be rational. It is to this alternative challenge that I turn in closing.

**3. The Improper Use Challenge**

The second challenge to claims of individual rationality from experimental evidence holds that agents do arrive at decisions or judgements using the mechanisms of classical rationality but their use of these reasoning systems fails (in some systematic way) to reach the standards required for rational decision making.[[21]](#footnote-21) The evidence for this stance comes, it seems, from three main sources: the Wason Selection Task, biased evidence gathering or assimilation (‘motivated reasoning’), and what is termed ‘belief polarization due to belief disconfirmation’.

**3.1 Wason Selection Task[[22]](#footnote-22)**

In the classic task designed by Peter Wason, subjects are given a standard piece of conditional reasoning: they are told that there is a deck of cards where each card has a number on one side and a letter on the other. They are then shown a display revealing one side of four different cards, as in Figure 1.

**E**

**7**

**2**

**F**

*Figure 1: Wason Selection Task cards*

Subjects are then asked:

* Which card(s) must be turned over to test the rule that if a card has a vowel on one side then it has an even number on the other side?

What Wason 1966 found was that only c.10% of participants deliver the logically correct answer (which is ‘E’ and ‘7’ in the above, as only the reverse of these cards could falsify the rule). Many participants, for instance, incorrectly think the ‘2’ card needs to be turned over. However, turning over the ‘2’ card is irrelevant as, if the ‘2’ card has a vowel on the reverse, that fits with the rule, and if it has a non-vowel on the other side that doesn’t matter (as the rule doesn’t say “all *and only* vowels have even numbers on the other side”). As Chater & Oaksford 1999: 62 summarise:

According to Popper’s falsificationist view of science, it is never possible to confirm a rule, only to disconfirm it. The rule will be disconfirmed only if there is a logical contradiction between the hypothesis and the data; that is, this requires finding a *p*, *not*-*q* card, which is explicitly disallowed by the rule. Hence the *p* card should be chosen, because it might have a *not*-*q* on the back, and the *not*-*q* card should be chosen, because it might have a *p* on the back, and no other cards should be chosen. This is the ‘logical’ solution to the task...The fact that people do not follow this pattern has been taken as casting doubt on human rationality.

***3.2 Myside Bias and Motivated Reasoning***

It turns out that subjects are prone to a range of mistakes around their search for and incorporation of evidence. They (i) hang on to beliefs in light of counter-evidence or don’t search properly for counter-evidence (‘belief preservation’), (ii) are overconfident (attributing too high a credence to extant beliefs), (iii) are subject to belief polarization (where groups of like-minded individuals reinforce a shared belief via poor reasons). So, for instance, a subject who has already decided to buy a Honda rather than a Toyota will look longer at Honda ads and skip over Toyota ones (Mandelbaum 2018), revealing that the subject’s prior beliefs improperly influence their search for evidence, leading them to seek out evidence that supports what they already believe and ignore or discount counter-evidence. This effect has been widely demonstrated in the experimental literature, e.g. subjects who antecedently believe in the efficacy of capital punishment hear equivocal evidence as confirming the truth of this belief, while those who antecedently reject the death penalty hear the same evidence as supporting their belief (Lord, Ross, & Lepper 1979), while smokers turn out to be more receptive to arguments denying a link between smoking and lung cancer than non-smokers are (Brock & Balloun 1967). These flaws count as instances of what Mercier & Sperber 2018: 244 call the ‘Myside Bias’ (which connects to the Confirmation Bias noted in §2). The existence of the Myside Bias spells trouble for claims of individual rationality, since:

Not only does reasoning fail to fix mistaken intuitions, as this approach claims it should, but it makes people sure that they are right, whether they are right or wrong, and stick to their beliefs for no good reason. Historical examples attest that these are not minor quirks magnified by clever experiments, but real phenomena with tragic consequences.

While our biased treatment of evidence means, Mercier and Sperber argue, that we must reject the idea that *individuals* are rational, it is nevertheless, they suggest, an adaptive feature for it underpins the key role that reasoning plays in arriving at better *group* decisions. [[23]](#footnote-23) Since everyone somewhat blindly advocates for their own extant beliefs, the group as a whole is exposed to the strongest possible arguments for each side. Myside bias, Mercier and Sperber 2018: 331 argue, thus reveals that reason is fundamentally a social competence and its “bias and laziness aren’t flaws; they are features that help reason fulfil its function”.

***3.3 Protecting an individual’s sense of self***

Mandelbaum 2018 is concerned with the peculiar phenomenon whereby sometimes subjects appear to *increase* their level of belief in a given proposition in the face of *accepted* disconfirming evidence (‘belief polarization due to belief disconfirmation’). Yet this is the exact opposite of what is predicted by a Bayesian belief-updating procedure. The clearest example of this phenomenon occurs with respect to religious beliefs (Batson 1975) and a particularly noteworthy case involves millennial cults, where members sometimes increase their belief in the truth of the cult’s proclamations *after* the date on which the world was previously predicted to end (Dawson 1999). As Mandelbaum stresses, however, the phenomenon also apparently occurs with more mundane beliefs and he cites experimental work concerning things like an individual’s attitude to the safety of nuclear power (Plous 1991) or affirmative action or gun control (Taber & Lodge 2006). These kinds of cases – where, as Mandelbaum argues, subjects seem to ditch Bayesian principles in order to protect their sense of self, hanging on to beliefs that shape their understanding of themselves – demonstrate that the systems of belief updating we use are systematically designed in a way that can yield non-rational decisions (see also Kahan 2017). That is to say, the system is “set up to properly output the actions we categorize as irrational” (Mandelbaum 2018: 4).

 So, should we accept that experimental evidence demonstrates our use of logical decision-making systems is systematically flawed?

**3.4 Challenging the evidence for flawed use of logical systems**

A first point to note is that standard answers in the Wason Test are only wrong/irrational *if* subjects are reasoning using Popperian (rather than Bayesian) methods. That is to say, turning over the red card in our above example, while it can’t falsify the rule, could provide further evidence capable of raising an individual’s credence in the rule. So, from a Bayesian perspective, turning over the red card might count as acceptable (Oaksford & Chater 1994). Furthermore, as noted in §2.3, individuals tend to do better in reasoning tasks if ecological validity is increased, and, as Cosmides & Tooby 1992 have shown, this is certainly the case with the Wason Task.[[24]](#footnote-24) So the idea that the Wason Task really challenges claims of individual rationality is, I think, moot (Stenning & van Lambalgen 2008).

 More importantly, however, I want to take issue with the assumption behind the argument from Mercier & Sperber (and at play in many other places) that motivated reasoning (where what one already believes shapes one’s search for or incorporation of evidence) is not, generally speaking, individually rational. Since we have already accepted (§2.4) that rational processes may look at *some* rather than *all* evidence and arrive at *good enough* (rather than *optimal*) decisions, a process which ignores some counter-evidence can nevertheless, it seems, count as entirely rational.[[25]](#footnote-25) Claiming that motivated reasoning is always irrational would ignore the fact that people usually have reasons for believing what they do (e.g. evidence from testimony or perception), and that adopting new hypotheses is itself not a cost-free exercise (requiring cognitive expenditure in terms of revising previous sets of beliefs, etc). So, the idea that there should be a degree of ‘stickiness’ about the beliefs one already holds, or that a subject should restrict how much evidence she looks for, far from showing that individual rationality does not hold, seems to be something we might *predict* given the assumption that individuals are, broadly speaking, rational.[[26]](#footnote-26) Skipping Toyota ads once one has made up one’s mind to buy a Honda, contra the claims above, can be rational, as it allows one to shift scarce cognitive reserves to other decision-making issues.

 Of course, there is a balance to be struck here: too great a degree of belief preservation (i.e. scepticism about presented evidence) *can* slide over into irrational fact-blindness/unwarranted dismissal of counter-evidence, and work needs to be done in any particular case to discover where the line should be drawn (Kahan 2017). Yet in principle, I suggest, motivated reasoning can be rational. And in fact, this point is already recognised within the experimental work cited in support of this second challenge to individual rationality. Thus Taber & Lodge 1992: 768 write:

How we determine the boundary line between rational skepticism and irrational bias is a critical normative question, but one that empirical research may not be able to address. Research can explore the conditions under which persuasion occurs (as social psychologists have for decades), but it cannot establish the conditions under which it should occur. It is, of course, the latter question that needs answering if we are to resolve the controversy over the rationality of motivated reasoning.

This is to recognise that the mere existence of motivated reasoning is not on its own sufficient to challenge claims of individual rationality.

 Finally, then, turning to belief polarization due to belief disconfirmation, I suggest that the evidence for this phenomenon with respect to more ‘ordinary’ beliefs (like belief in the safety of nuclear power, etc) is in fact fairly weak. For the experimental work in these cases seems better treated as examples of motivated cognition. Someone who antecedently believes in the safety of nuclear power may well be more likely than others to interpret an ambiguous or equivocal piece of evidence in a way that supports their extant belief (Plous 1991), but in and of itself, as argued above, this is not an objection to individual rationality. What would need to be shown in addition is that a given individual has strayed across the line from warranted belief preservation into irrational fact blindness. Yet, if this can be shown, then a supporter of individual rationality, no less than a sceptic, might be happy to treat this individual as failing to be rational about this issue. For the claim that individuals generally act in line with what they believe, desire, etc, is supposed to be a claim about *typical* intentional action, rather than a universal claim about all decisions. Clearly, there will be exceptions. Relatedly, although Mandelbaum may well be right that for certain kinds of beliefs (e.g. religious beliefs) Bayesian reasoning does indeed break down (in favour of some kind of self-protection), still it is unclear whether this is sufficient to show assumptions of individual rationality mistaken (Van Leeuwen 2014). An advocate of individual rationality must already allow that there are a range of cases where a further story needs to be told, e.g. allowing for akrasia, cases of conflicting occurrent and dispositional beliefs, and episodes of mental disorder (e.g. see Egan 2008 on compartmentalized beliefs). That the domain of non-rational thinking may extend to matters of *faith,* like religious beliefs, seems to me to be a conclusion that a supporter of individual rationality could well live with.

**4. Conclusion**

The claim that, typically, our actions are the result of rational processing of our reasons has been challenged in recent years by a range of experimental evidence which seems to show either that we commonly make use of a non-logical/automatic decision-making system or that our use of our logical systems is systematically flawed. However, I have tried to suggest here that the rejection of individual rationality on the basis of this experimental evidence may well be too swift. Contra advocates of the Automatic System I think we should reject much of the evidence for irrational decision-making (explaining it through greater sensitivity to pragmatics). I also suggested that we should reject the entire Dual Process model on which this first challenge is based. Contra the second challenge, I argued that belief preservation or motivated reasoning is (at least in principle) rational. Only belief polarization due to belief disconfirmation is truly problematic, but these are special cases where claims of rationality may indeed be tenuous. I conclude, then, that experimental evidence should not (currently) be taken to refute claims of individual rationality.

**References**

Ariely, D. (2008). *Predictably Irrational*: Harper.

Batson, C. D. (1975). Rational processing or rationalization? The effect of disconfirming information on a stated religious belief. *Journal of Personality and Social Psychology, 32*, 176–184.

Borg, E. (manuscript). *Compelling Reasons.*

Carruthers, P. (2017). Mindreading in adults: evaluating two-systems views. *Synthese, 194*(3), 673-688.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends Cogn Sci, 3*(2), 57-65.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptions for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford: Oxford University Press.

Davidson, D. (1963). Actions, Reasons, and Causes. *The Journal of Philosophy, 60*(23), 685-700.

Dawson, L. (1999). When prophecy fails and faith persists: A theoretical overview. *Nova Religio, 31*, 60-82.

Dutilh Novaes, C. (2018). The enduring enigma of reason. *Mind and Language, 33*, 513–524.

Egan, A. (2008). Seeing and believing: perception, belief formation and the divided mind. . *Philosophical Studies, 140*, 47–63.

Evans, J. (2009). How many dual-process theories do we need? One, two, or many? In K. Frankish & J. Evans (Eds.), *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press.

Evans, J. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning, 18*(1), 5-31.

Evans, J., & Stanovich, K. E. (2013). Dual-Process theories of higher cognition. *Perspectives on Psychological Science, 8*, 223-241.

Fisher, S. A., & Mandel, D. R. (2021). Risky-choice framing and rational decision-making. *Philosophy Compass, 16*(8), e12763.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of economic perspectives, 19*(4), 25-42.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*(62), 451-482.

Hertwig, R., & Gigerenzer, G. (1999). The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12*(4), 275-305.

Kahan, D. (2017). Misconceptions, misinformation, and the logic of identity-protective cognition. *Social Science Research Network*.

Kahneman, D. (2011). *Thinking fast and slow.* London: Penguin.

Karlan, B. (2021). Reasoning with heuristics. *Ratio, 34*(2), 100-108.

Kripke, S. A. (2011). *Philosophical Troubles: Collected Papers*. Oxford: Oxford University Press.

Kruschke, J. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior, 36* 210-226.

Kühberger, A. (2002). The rationality of risky decisions. *Theory and Psychology, 12*(4), 427–452.

Lieder, F., & Griffiths, T. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences, 43*, 1-60.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*(11), 2098-2109.

Mandelbaum, E. (2018). Troubles with ayesianism: an introduction to the psychological immune system. *Mind and Language, 34*, 1-17.

Mercier, H., & Sperber, D. (2018). *The Enigma of Reason: A New Theory of Human Understanding*. London: Penguin.

Mosconi, G., & Macchi, L. (2001). The role of pragmatic rules in the conjunction fallacy. *Mind & Society 2*, 31-57.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608–631.

Plous, S. (1991). Biases in the assimilation of technological breakdowns: Do accidents make us safer? . *Journal of Applied Social Psychology*(21), 1058–1082.

Samuels, R., Stich, S., & Bishop, M. (2002). Ending the rationality wars: how to make disputes about human rationality disappear. In R. Elio (Ed.), *Common Sense, Reasoning and Rationality* (Vol. Vancouver Studies in Cognitive Science, Vol. 11). New York: Oxford University Press.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*

*119*, 3–22.

Stein, E. (1996). *Without Good Reason.* Oxford: Clarendon Press.

Stenning, K., & van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. Cambridge, MA, US: MIT Press.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science, 50*(3), 755-769.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory and Cognition, 39*, 1275–1289.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General, 143*(5), 1958-1971.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458.

Van Leeuwen, N. (2014). Religious credence is not factual belief. *Cognition, 133*(3), 698-715.

Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New Horizons in Psychology.* . Harmondsworth: Penguin.

1. \* Thanks are due to Nat Hansen and Sarah Fisher, and audiences at St. Andrews and Reading. [↑](#footnote-ref-1)
2. See, e.g., Davidson 1963. [↑](#footnote-ref-2)
3. For discussion see, e.g., Lieder & Griffiths 2020, Kühberger 2002, Stein 1996. [↑](#footnote-ref-3)
4. Doris 2015: 64-6 frames this challenge as follows: “Where the causes of [an agent’s] cognition or behaviour would not be recognised by the actor as reasons for that cognition or behaviour, were she aware of these causes at the time of performance, these causes are *defeaters.* Where defeaters obtain, the exercise of agency does not obtain. If the presence of defeaters cannot be confidently ruled out for a particular behaviour, it is not justified to attribute the actor an exercise of agency”. The problem is that (2015: 68) “The empirical evidence indicates that defeaters occur quite frequently in everyday life”. [↑](#footnote-ref-4)
5. This paper focuses on the challenge to individual rationality, paradigmatically as it emerges from Kahneman et al’s ‘heuristics and biases programme’. As is well-known, however, alternative views (e.g. Gerd Gigerenzer et al’s ‘adaptive rationality’ model) characterise heuristic use as rational. Although beyond the scope of this discussion, I’d argue that the objection raised in §2.4, which holds that properties like speed and frugality don’t define types of cognition, also challenge alternative approaches like Gigerenzer’s. [↑](#footnote-ref-5)
6. For a somewhat fuller coverage of the evidence, see Borg (manuscript). [↑](#footnote-ref-6)
7. Although there are important distinctions between ‘system’ and ‘process’ talk (see Evans 2009), I’ll skate over the issue here, using the two terms interchangeably. [↑](#footnote-ref-7)
8. Presenting heuristic thinking via comparison to visual judgements is common. For instance, Thaler and Sunstein 2008: 19-21 open their discussion with an optical illusion concerning relative size, commenting that “[Y]our judgement in this task was biased, and predictably so…Not only were you wrong; you were probably confident that you were right”. See also Mercier and Sperber’s 2018 introduction to heuristic thinking. [↑](#footnote-ref-8)
9. As Toplak, West, & Stanovich 2011: 1275-6 note “The three problems on the CRT are of interest to researchers working in the heuristics-and-biases tradition because a strong alternative response is initially primed and then must be overridden. As Kahneman and Frederick made clear…this framework of an incorrectly primed initial response that must be overridden fits in nicely with currently popular dual process frameworks”. [↑](#footnote-ref-9)
10. While the original CRT contained just three questions something of a cottage industry has grown up generating CRT-like questions; see, e.g., Trouche, Sander, & Mercier 2014, Thomson and Oppenheimer 2016. [↑](#footnote-ref-10)
11. Lieder & Griffiths 2020: 3 note a potential worry with this proliferation: “Unfortunately, the number of heuristics that have been proposed is so high that it is often difficult to predict which heuristic people will use in a novel situation and what the results will be”. [↑](#footnote-ref-11)
12. The example follows the general shape of the infamous ‘Asian flu’ example from Tversky & Kahneman 1981. [↑](#footnote-ref-12)
13. For instance, concerns have been raised about replication for Danziger et al’s hugely influential 2011 claim that hungry judges give stiffer penalties (e.g. Hagger et al. 2016, Weinshall-Margel & Shapard, 2011), while John Bargh’s finding that people move more slowly after exposure to prompts about old age has failed numerous attempted replications. See the ‘Many Labs: Investigating variation in replicability’ project, https://osf.io/wx7ck/ [↑](#footnote-ref-13)
14. See, for instance, Ho and Imai 2008 on the so-called ‘ballot order effect’. See also §3.4 below. [↑](#footnote-ref-14)
15. This is to agree with Glöckner 2016: 610 that “sometimes there is a nonobvious rational basis for irrational-looking behavior”. [↑](#footnote-ref-15)
16. Discussion here has benefitted from conversations with Sarah Fisher. [↑](#footnote-ref-16)
17. Samuels, Stich, & Bishop 2002 also raise this objection, which they label the ‘cross-over problem’. See also Evans 2012, Carruthers 2017. [↑](#footnote-ref-17)
18. E.g. Kahneman 2011: 20: “System 1 operates automatically and quickly, with little or no effort”. [↑](#footnote-ref-18)
19. This point has been recognised by advocates of Dual Process models recently and it has led some (e.g. De Neys 2018) to claim that the right approach is a ‘hybrid’ one, where the Automatic System is capable of delivering *both* intuitive and (simple) logical answers. However, it is very unclear that this kind of model can do the explanatory work the Dual Process approach was originally introduced to do (e.g. explaining the apparent prevalence of non-logical answers in experimental cases), nor whether, once this concession is made, much of substance really remains of the Dual Process approach. [↑](#footnote-ref-19)
20. In line with Evans and Over 1996, Karlan 2021 distinguishes ‘locally rational solutions’ (good enough solutions for the problem in context) from ‘ideally rational solutions’ (modally robust results, in line with Classical Rational Choice), arguing that both notions of rationality are needed. [↑](#footnote-ref-20)
21. As Mandelbaum 2018: 4 notes, what is needed to show a Rational Choice approach false “is to find actions that are not just the result of errors in processing. Rather, the irrationality has to result from a system that is set up to properly output the actions we categorize as irrational”. [↑](#footnote-ref-21)
22. The Wason Task is usually placed alongside the CRT, providing evidence for the Automatic System. However, I think it is better viewed as supporting an Improper Use challenge. For in the Wason case participants do not have an immediate, intuitive feel for an answer, nor are they confident in their responses. Instead, participants think careful and deliberatively but still make mistakes. [↑](#footnote-ref-22)
23. Cf Dutilh Novaes 2018. [↑](#footnote-ref-23)
24. Asked to turn over cards relating to the rule ‘You must be over 21 to drink alcohol’, participants were far more likely to turn over the potentially falsifying cards, e.g. showing ‘beer’ or ‘18’. [↑](#footnote-ref-24)
25. Mandelbaum 2018 notes that some Bayesian models actually *predict* biased assimilation. [↑](#footnote-ref-25)
26. For ‘stickiness’ see Kripke 2011: 49, Levy 2022. [↑](#footnote-ref-26)