Nonrational Belief Paradoxes as Byzantine Failures

0. Abstract

David Christensen and others argue that Dutch Strategies are more like peer disagreements than Dutch Books, and should not count against agents' conformity to ideal rationality. I review these arguments, then show that Dutch Books, Dutch Strategies, and peer disagreements are only possible in the case of what computer scientists call Byzantine Failures—uncorrected Byzantine Faults which update arbitrary values. Yet such Byzantine Failures make agents equally vulnerable to all three kinds of epistemic inconsistencies, so there is no principled basis for claiming that only avoidance of true Dutch Books characterizes ideally rational agents. Agents without Byzantine Failures can be ideally rational in a very strong sense, but are not normative for humans.

Keywords: Dutch books, Dutch strategies, Reflection, ideal rational agents, Byzantine generals, peer disagreements

1. Consistency Paradoxes for Ideal Rational Agents

1.1 Paradoxes of Rational Requirements

The following characteristics are often taken to characterize an ideally rational agent (Grüne-Yanoff 2007):

1.1.1 the agent's preference ordering over her prospects[1] is complete

1.1.2 the agent's preference ordering over her prospects is transitive

[1] The set of prospects at any time is fixed, and each prospect is either a future state of the world which occurs with certainty or a probability distribution over such states.

1.1.3 the agent's preference ordering over her prospects is continuous

1.1.4 the agent's preference ordering over her prospects is independent of irrelevant alternatives

1.2.1 the agent's set of probabilistic beliefs is coherent (they satisfy the Kolmogorov axioms)

1.2.2 the agent's set of probabilistic beliefs is complete

1.2.3 the agent updates her probabilistic beliefs by conditionalization

Frank Ramsey and Bruno de Finetti discovered a natural way of unifying these perhaps seemingly disparate characteristics through the phenomenon of Dutch Books. In a Dutch Book, a bettor faces a guaranteed loss (regardless of the outcome of any risks hazarded), when making a series of synchronic bets at her fair betting quotient[2] against a competent bookie who possesses no evidence not also in the possession of the bettor (Vineberg 2016). Characteristics 1.1.1-1.1.4 are the axioms of von Neumann and Morgenstern's Expected Utility Theory (1953), which gives the standard method for assigning value under conditions of risk, and hence for interpreting the notion of a guaranteed loss. Characteristic 1.2.2 ensures that the better actually has a fair betting quotient for all of the bets offered by the bookie. Ramsey (1964) and de Finetti (1964) then show that unless the bettor possesses characteristic 1.2.1, she can face a Dutch Book. While the pragmatic connections among guaranteed losses, optimal bets, and ideal rationality are perhaps tenuous and difficult to define, the possibility of a Dutch Book is nonetheless a plausible illustration of a failure of ideal rationality (Skyrms 1987). When the series of bets is offered diachronically, a guaranteed-loss situation is called a Dutch Strategy,[3] which Teller (1973) and Armendt (1980) show results for any agent lacking

[2] A fair betting quotient is the odds at which the bettor is equally willing to take either side of the bet.

[3] Skyrms (1993) gives the exact conditions for such a diachronic series.

characteristic 1.2.3. Since Dutch Books and Strategies connect the Expected Utility Theory axioms, the Kolmogorov axioms for probability theory, and Bayesian reasoning—each of which has been enormously fruitful—they seem to have explanatory power for characterizing ideally rational agents. The characteristics they demand can be summed up as "epistemic consistency" (Christensen 1991).

Bas van Fraassen (1984) and Jordan Sobel (1987) show that avoiding Dutch Strategies also justifies another proposed characteristic of ideally rational agents: Reflection. The principle of Reflection demands strong diachronic consistency in judgments, such that "the agent's present subjective probability for proposition A, on the supposition that his subjective probability for this proposition will equal r at some later time, must equal this same number r" (van Fraassen 1984).[4] David Christensen (1991) worries that Reflection leads to paradoxes—most seriously a contradictions with the Kolmogorov axioms in a situation where an agent has a small-but-non-zero credence that she will in the future have credence .95 that she has no credences greater than .90.[5] Reflection means that an agent who will be irrational in the future must be irrational today, a result Christensen takes as absurd. W.J. Talbott (1991) improves on Christensen's argument in two regards. First, he shows that

[4] In other words, a change in credence requires a change in evidence. Credences of ideally rational agents, like stock prices in efficient markets, must "already reflect the effects of information based both on events that have already occurred and on events which, as of now, the market expects to take place in the future…the full effects of new information on intrinsic values [will] be reflected 'instantaneously' in actual prices…[so]…successive price changes in individual securities will be independent...[and]…the future path of the price level of a security is no more predictable than the path of a series of cumulated random numbers" (Fama 1965). In fact, because prediction market prices can be interpreted as credences (Wolfers and Zitzewitz 2006), the theory of efficient markets (where traders get no free lunch) and ideally rational agents (where bookies get no free lunch) have the same constraints.

[5] Perhaps because a typically reliable informant has informed her that her drink was spiked with the drug LSP which has this unusual psychedelic effect, though in this case the informant erred. Such cases implicate not just Reflection but also deductive closure of justification (Backes 2019).

the general formula for generating Christensen cases is any situation in which an agent expects that she will violate Conditionalization (characteristic 1.2.3). Second, he gives an everyday example in which an agent expects that she will violate Conditionalization without doing anything obviously irrational: all the agent has to do is (1) have credence r about the contents of her breakfast on day T (today) and (2) expect that on day T+365 she will have a credence less than r about the contents of her breakfast on day T. We are clearly all ineluctably vulnerable to Dutch Strategies.

1.2 Equivalence of Single-Agent Diachronic Consistency and Two-Agent Synchronic Consistency

Christensen (1991) shows that single-agent Dutch Strategies are equivalent to Double Agent Dutch Books. In a Double Agent Dutch Book, a bookie makes a sure profit on a set of synchronic bets with a pair of bettors whose credences differ. We can easily convert any Dutch Strategy into a Double Agent Dutch Book by simply replacing the future agent in the description with a parallel agent. If the parallel agents' prospects are entangled (e.g. by joint finances), then the bookie's sure gain implies a sure loss for both of them. In a further (unnamed) variation, which Christensen discusses as an inconsistency without actually giving a Dutch Book, the agents' credences need not actually differ as long as one agent believes that they differ. If I am willing to bet 3:1 odds-on that Reflection is a true characterization of all rational agents and also willing to bet 3:1 odds-on that Christensen will bet odds-against this claim, then the bookie makes a sure profit no matter whether Christensen (having come around) prefers 3:1 odds-on for Reflection or (still holding out) 3:1 odds-against Reflection. The bookie's payoffs are given in Table 1 (when she varies her stakes as indicated there).

Table 1

	Reflection is True and	Reflection is True and	Reflection is False and	Reflection is False and

	Christensen bets 3:1 odds-on	Christensen bets 3:1 odds-against	Christensen bets 3:1 odds-on	Christensen bets 3:1 odds-against
My bet on Reflection (7x stake)	-7	-7	21	21
My bet on Christensen's bet on Reflection (5x stake)	15	-5	5	-15
Christensen's bet on Reflection (5x stake)	-5	15	-5	-5
Total	3	3	21	1

The bookie has developed a Double Agent Dutch Book just by knowing that I think I disagree with Christensen. In a way this is unsurprising: Dutch Books are tests of epistemic consistency, and peer disagreement seems like it can be characterized as group inconsistency.[6] Christensen, however, stresses that such group inconsistency is not indicative of any failure of ideal rationality in the agents who make up the group—perhaps, for instance, the agents have reasonably differing priors.

1.3 Limitations on Expectations of Consistency in Ideal Rational Agents

Christensen (1991) argues that since Dutch Strategies lead to paradoxes and their structurally-identical Double Agent Dutch Books do not indicate failures of ideal rationality, Dutch Strategies themselves should not be interpreted as constraints on ideally rational agents. This nonetheless comes at a cost for Christensen, since such Dutch Strategies are the leading support for Conditionalization (characteristic 1.2.3) which Christensen accepts. Since Talbott (1991)'s examples show that humans cannot always expect to obey Conditionalization (yet he thinks we ought to be rational and ought-implies-can), he jettisons

[6] This point has been formalized much earlier by Ryder (1981).

that principle along with Reflection and Dutch Strategy avoidance in general. Talbott takes it that only Dutch Books and Strategies where the agent is aware of the guaranteed loss constrain rationality, but this renders them fruitless as tests of general epistemic consistency. Surely rationality requires more than avoiding explicit guaranteed losses.

Christensen himself later brings pressure from two directions against this approach of relaxing constraints on ideal rationality. First, he treats peer disagreement as a source of epistemic concern for rational agents (Christensen 2000; 2007b). Second, in the presence of irrational beliefs even purely Synchronic Reflection also leads to paradoxes, even though it is supported by a simple single agent Dutch Book (Christensen 2007a). Christensen releases this pressure by weakening the constraints yet further: we shouldn't expect perfect synchronic meta-consistency, either (2007a). The arguments for it aren't a true Dutch Book, Christensen says, because the bookie has *contingent* knowledge that the agent doesn't have—it just happens to be knowledge about the agent's own credences (Christensen 2007a). Credences—whether synchronic or diachronic, first-party or third-party—are just ordinary evidence (Christensen 2007a). Sherrilyn Roush (2009) uses the idea that credences are just ordinary evidence to develop a Re-Cal variant of Conditionalization for rational updating of credences even in the face of first-order Conditionalization failures. Because this method relies on principled distinctions between first-, second-, and higher-order evidence, credences, and Conditionalization, it is of no assistance for resolving cases where the non-rational first-order credences are not governed by higher-order credences and thus subject to revision. Peer disagreement is just a special case of this latter situation: neither of the peers' credences are higher-order with respect to the other, so there is no rational way to resolve the incoherence (Roush 2009).

These arguments naturally lead to a three-fold categorization of epistemic consistency demands: strict constraints on rationality supported by true Dutch Books, broader principles

supported by Dutch Strategies that should be used when reality doesn't conspire against us (Vineberg 1997), and cases of pure inconsistency lacking any principled method for resolution. Ideally rational agents should be untroubled by peer disagreement, avoid Dutch Strategies whenever they can do so without paradox, and avoid Dutch Books at all costs. Only vulnerability to true Dutch Books should worry us concerning an agent's characterization as ideally rational.

2. The Byzantine Failure Explanation of Consistency Paradoxes

2.1 Byzantine Generals and Byzantine Failures in Computer Science

The large philosophical literature generating and analyzing the paradoxes that result when supposedly ideal rational agents are confronted with nonrational beliefs can be understood as instances of what computer scientists call the Byzantine Generals problem. The thought experiment given by Lamport, Shostak, and Pease (1982) runs as follows. A number of generals from Byzantium are encamped around a city they have under siege, each with his own army. They are trying to decide whether to storm the city or retreat until the next campaign season, but face the difficulty that some of their number may be traitorous. The constraints on their decision-making are that all loyal generals must adopt the same plan (lest their forces be scattered and routed) and that plan must be the one that a majority of loyal generals privately think best (lest the traitors control the army's strategic decision-making to their advantage).[7] Under what conditions can these constraints be met? Given Kenneth May (1952)'s theorem in favor of simple-majority voting for two-candidate ballots, a first instinct

[7] One may note a certain analogy to Kenneth Arrow (1950)'s impossibility theorem for converting individual ordinal preferences to community ordinal preferences under conditions of unrestricted domain, non-dictatorship, Pareto efficiency, and independence of irrelevant alternatives. Decision theory has already been analyzed in these terms by Briggs (2010). In the Byzantine Generals case, the domain has been restricted, but the non-dictatorship requirement has been strengthened.

is to assume that the constraints are met as long as the super-majority among loyal generals is greater than the number of traitors. The trouble is that in the Byzantine scenario there is no neutral arbiter to count the ballots, and a traitorous general may send different responses to different loyal generals in order to sow disarray.

Lamport et al. (1982) derive three important results from the Byzantine Generals problem. The first is that it is equivalent to the Byzantine Lieutenants problem, wherein all loyal Lieutenant Generals adopt the same plan, and it is the plan ordered by the Field Marshal as long as the Field Marshal is loyal. Hierarchy in place of anonymity provides no assistance if the hierarchy cannot be trusted. The second result is that the problem cannot be solved without $3t + 1$ generals, where t is the number of traitors. The third result is that if traitors can be caught when forging messages (e.g. by enforcing cryptographic signing), then the naïve supermajority solution holds, because each general can report every message he receives to every other general without possibility of deception.

While the canonical form of the Byzantine Generals problem involves malicious actors, Lamport et al. (1982) are clear that it applies just as strongly to ordinary hardware failures which result in different signals being received by different processors. In fact their earlier more rigorous and less didactic paper (Pease, Shostak, and Lamport 1980) mentions only faulty processors and not traitorous generals. Here the constraint is merely that "independent processes" must "arrive at an exact mutual agreement of some kind" (Pease, Shostak, and Lamport 1980). A system which meets this constraint exhibits "interactive consistency" (Pease, Shostak, and Lamport 1980). A faulty processor can play the role of a traitorous general merely by reporting different values to different peer processors. When two processors disagree about the value of an input, this is merely the Lieutenants version of the problem (Lamport, Shostak, and Pease 1982). Further, "processor" means nothing more than a peer agent in a parallel system (Lamport, Shostak, and Pease 1982) or even a subsequent

independent state of a single system (Biely and Hutle 2009). Later papers on the Byzantine Generals problem thus often recast it in terms of "Byzantine Faults" which "present different symptoms to different observers" and "Byzantine Failures" in which systems requiring interactive consistency cannot achieve it due to Byzantine Faults (Driscoll et al. 2004). If a Byzantine Fault is detected and corrected, whether by a trusted meta-process or a robust consensus protocol, then it will not result in a Byzantine Failure (Arora and Kulkarni 1998).[8] Since arbitrary hardware failures lead to arbitrary processing results, any arbitrary hardware failure can easily lead to a Byzantine Fault (Lamport, Shostak, and Pease 1982; Driscoll et al. 2004). This leads Arora and Kulkami (1998) to simply define Byzantine Faults as those which "corrupt processes permanently[9] and undetectably[10] such that the corrupted processes execute arbitrarily nondeterministic[11] actions." Such processes will obviously be inconsistent with the correctly-functioning processes. Biely and Hutle (2009) call Byzantine Faults "arbitrary value faults" because the result is that there is no constraint on the output value of the process. Byzantine Faults are the most general model of faults because they do not assume that any degree of detection and correction is possible (Biely and Hutle 2009).

2.2 Peer Disagreement Cases as Byzantine Failures

Peer disagreement cases are the most obvious instances of Byzantine Failure in human agents. In the check-splitting case (Christensen 2007b), two peers need to come to

[8] Kuznets et al. (2019) provide an epistemic logic for checking whether Byzantine Faults can be caught.

[9] I have left out the complicated discussion of timing in the Byzantine Generals literature because unlike real carbon or silicon agents, Dutch Strategies operate on a turn-based system. Permanent in this context merely means extending beyond the time-out in a real-time system or until the end of the turn in a turn-based system.

[10] Undetectable by the system itself, because if a process detects its own fault, then it will not report it, whereas if a neutral arbiter does so, then that process is no longer a peer. This does not mean that the fault is undetectable in principle by an arbiter outside the system.

[11] Arbitrary and nondeterministic not in the strong sense of appealing to irreducible objective chance but in the sense that the result cannot be predicted by knowing the algorithm used by the processor.

consensus about the total bill so that each pays the correct amount. The peers produce inconsistent answers. If each interpreted a smudged line on the bill differently, we have the faulty-input Byzantine Lieutenants problem. Since both know how to perform arithmetic, if one has added incorrectly then it is due to an arbitrary, non-deterministic fault like skipping a line, adding a line twice, failing to carry, etc. The agent did not catch this fault before making her report. There is no detector available (e.g. a trusted third party, or a checksum algorithm). It does not matter whether the error leads to forged responses or not,[12] because there are not enough agents available to perform even the naïve majoritarian consensus protocol. The Byzantine Fault has led to a Byzantine Failure where there is no correct procedure for achieving consensus—the system lacks interactive consistency.

Analysis of the check-splitting case in more traditional terms yields the same result. If both agents stand fast then there is a Double Agent Dutch Book against them—they are epistemically inconsistent. The parties can take each other's credences as evidence and use Conditionalization to update their own credences, but doing so won't generally result in convergence since their priors differ. In fact, it can lead to paradoxical situations where credences cross over (Lang 2014). Meta-methods like Re-Cal won't work because the situation is symmetric (Roush 2009). The parties can merely decide to split the difference, but now they are assuming that both have made errors rather than only one, and that those errors are precisely canceling—a highly unlikely set of events, for which there is no evidence. If that were a rational requirement, then rationality would be anti-truth-conducive. In short, the agents are stuck in a situation of epistemic inconsistency without any generalizable and reliable means of escape.

[12] As Driscoll et al. (2004) make clear for the silicon case, this should not be taken for granted as it often is. If a hardware error can make a person calculating a total read a line incorrectly while doing the sum, could not the same or similar error make a person read the line incorrectly while reporting the results of her calculation?

The other Double Agent Dutch Book cases Christensen (1991) discusses are relevantly similar. He portrays himself as holding a trusted meta-role when he explains his wife's differing meteorological credences by her "pessimism," but unless she accepts him as a checker and corrector of her views rather than an epistemic peer, she has no reason to concede to that judgment. If she fails to concede to his judgment and holds fast to her credences, then a clever bookie can do guaranteed damage to their joint bank account. A narrator who accepts Christensen's view that she is unduly pessimistic will interpret her pessimism as an arbitrary hardware failure, where she fails to match her credences to the objective chances in accord with Lewis (1980)'s Principal Principle. Since there is a Dutch Strategy available in favor of the Principal Principle (Howson 1992), this serves to identify the agent experiencing the Byzantine Fault to third parties. What it does not do, given the unavailability of both a checker actually trusted by both parties and additional peer parties, is prevent the Byzantine Fault from leading to a Byzantine Failure where the parties exhibit interactive inconsistency.

Peers exhibit unresolvable epistemic inconsistency (vulnerability to a Double Agent Dutch Book) just in case they exhibit interactive inconsistency (Byzantine Failure). When agents exhibit interactive inconsistency, they have no reliable strategy available for achieving consensus, so they will be subject to Double Agent Dutch Books. When agents exhibit unresolvable epistemic inconsistency, they face guaranteed losses through Double Agent Dutch Books which both parties would wish to avoid if they had some reliable strategy available for achieving consensus.

2.3 Dutch Strategy Paradoxes as Byzantine Failures

As Christensen (1991) suggested, there is nothing fundamentally different about single-agent diachronic cases. Any Double Agent Dutch Book can be converted into a Dutch

Strategy by merely transferring the properties of the second agent to the first agent at a later time. If we expect time consistency from rational agents then this is a problem, otherwise not.

The same goes for the Byzantine Failure analysis of such cases. If I sum my own restaurant bill twice and get two different answers, I have an interactive inconsistency because the result should be the same and I have no more tools to resolve the failure than in the two-agent synchronic case. The agent who knows he will be unwarrantedly pessimistic in the future can only avoid treating the future self as a peer if the future self can be convinced that he is unduly pessimistic—but if the future self is aware of his pessimism and able to act on that knowledge then he can update using Roush's Re-Cal to escape the problem. If the future self is unconvinced of his own irrationality, then I am stuck treating him as a peer. If I assume that neither of us has experienced a Byzantine Fault, then he must have evidence that I lack and have updated his credences by Conditionalizing, so I should use Reflection to incorporate that information. If I assume that he has experienced a Byzantine Fault then I don't have a long enough time series (treating each temporal snapshot as a peer processor) to avoid Byzantine Failure. If I know that my undue pessimism will wear off, after all, then I can use Reflection to update directly to that post-pessimism correct value and there is no paradox.

Christensen (1991)'s catalog of psychological failures all amount to arbitrary hardware faults. In each case, the agent comes to believe something for some reason other than updating on evidence by Conditionalization, which is the rational algorithm that (as shown by Dutch Strategy) prevents diachronic epistemic inconsistency. In each case, the agent is unable to detect and correct his non-rational update. In each case, the resultant credence is essentially an arbitrary value. While less obvious, this is even true for Talbott (1991)'s forgetting case. When I forget what I had for breakfast, I have to update my credence, and I do not do so by Conditionalization on new evidence. What of Talbott's

ought-implies-can argument? In order to have a high credence in my choice of breakfast I need not remember the gestalt of consuming the breakfast—I need only store the credence from when I did remember the gestalt and refuse to update except by Conditionalization on new evidence. Characteristic 1.2.2 stated that ideal rational agents have a complete set of probabilistic beliefs—otherwise they might have no fair betting quotients for bookies to discover, be unwilling to take bets, and hence lack susceptibility to Dutch Books and Strategies not through rational success but rather through inadequacy. The agent with the fewest beliefs would be the most rational. If I have a complete set of probabilistic beliefs, however, then I must have adequate memory to store those, and cannot lose credences by memory pressure. If I lose credences and have to regenerate them from nearby credences (about e.g. what I usually have for breakfast), then I have experienced an arbitrary hardware failure. Surely Talbott is correct that this does not describe the human situation, in which such failures are inevitable, but it fails to do so in a way that is not unique to Dutch Strategies. In the other direction, we should expect arbitrary hardware faults to lead to vulnerability to Dutch Strategies. Memory faults do so, as Talbott showed. Computation faults would lead to incorrect Conditionalization—the only allowed update operation—which also results in a Dutch Strategy.

Christensen is therefore correct that not much separates Double Agent Dutch Book cases and Dutch Strategy cases. Not only are both subject to equivalent betting losses (assuming that consistency is demanded in the Double Agent case by e.g. entangled finances), but both are generated by Byzantine Faults. Both can be avoided by the same degree of enhanced redundancy.

2.4 Dutch Books as Byzantine Failures

Whereas Christensen draws a close analogy between Double Agent Dutch Books and Dutch Strategies, he distinguishes both sharply from true Dutch Books (1991; 2000; 2007a).

The latter he considers as genuine constraints on the credences of ideal rational agents. But what kind of irrationality is indicated by susceptibility to a Dutch Book? Brian Weatherson (2005) indicates that susceptibility to mathematical error is a sufficient kind of irrationality to make an agent vulnerable to Dutch Books. Prospects, after all, are probability distributions over payoffs. If you do the math wrong, you can easily find yourself in a Dutch Book.[13] And why might you do the math wrong? Well, you experienced an input, memory,[14] or calculation error that you didn't detect and correct: a Byzantine Fault. And as in the two-agent synchronic case, in the single-agent synchronic case every Byzantine Fault is trivially a Byzantine Failure. There is no justification for imputing some stronger form of irrationality to agents vulnerable to Dutch Books when math errors are both common and sufficient for such vulnerability. Conversely, every Byzantine Failure will lead to a Dutch Book. If the hardware failure isn't in credences—the arena subjected to a consistency demand by Dutch Books—then it isn't Byzantine. If the failure is in credences, then an arbitrary change to the credence for *p*, which leaves credences for *q*, *p & q*, etc. unaffected, will lead to a Dutch Book. Even explicit Dutch Books, of the type demanded by Talbott (1991), can be accepted in the event of Byzantine Failures: the fault need only erase the memory of the bookie presenting the guaranteed loss before accepting the series of bets.

Peer disagreement cases and Dutch Strategy paradoxes both presume Byzantine Failures. Unless there is an arbitrary value fault, there is no explanation for why the peers disagree or why the supposedly rational agent updates her credences other than by Conditionalization on new evidence. In fact, other human biases and limitations can be

[13] Weatherson (2004) argues that since Dutch Books only bind when consistency is expected, they do not mandate assigning a credence of 1 to all logical truths. Therefore there's no reason to assume that agents merely have credences rather than calculating them—certainly if humans can be ideal rational agents they would be the sort who sometimes have to calculate their credences.

[14] For the role of memory in deduction see Genot and Jacot (2020).

assimilated to Conditionalization by varying the payoffs, ensuring that in such non-Byzantine situations no Dutch Book is possible (Williams 2021). In the presence of Byzantine Failures, however, agents cannot guarantee that they will avoid Dutch Books. Agents can only satisfy ideal rationality if they can avoid Byzantine Failures—if, in Susan Vineberg (1997)'s phrasing, the universe declines to conspire against them.

Perhaps Christensen could respond that true Dutch Books test for epistemic consistency of states, rather than consistency of agents. Maybe the Dutch Book can only be offered while the putatively rational agent is in a constant state with respect to all her credences. Now, however, there can be no talk of bookies eliciting fair betting quotients—they must have direct access to the credences of the agent, and they must perform all the calculations with respect to the agent's preference ordering. The trouble with this approach is that states don't have preferences—agents do. Even more clearly, states do not experience payoffs. There is a reason that ideal rationality is an attribute of agents, rather than states.

3. Conclusion: A Stricter Model of Ideal Rationality

The conclusion is that, if ideal rationality is to mean anything at all, agents experiencing Byzantine Failures cannot count as ideally rational. In the absence of paradoxes generated by such failures, however, we have no reason to reject Dutch Strategy-motivated constraints on rationality. Such Dutch Strategies then provide a path to a stricter model of ideal rationality than that envisioned by Christensen and summarized in characteristics 1.1.1-1.2.3 at the start of this paper. The first characteristic which can be added is David Lewis's Principal Principle, supported by a Dutch Strategy given by Colin Howson (1992). Then, since the Principal Principle is incompatible with contingent priors (Milne 1991), another additional characteristic of ideally rational agents is that their priors will be necessary. Necessary a posteriori truths are discovered by evidence, so their priors would be necessary a priori. The most promising scheme for necessary a priori priors is Indifference (Pettigrew

2016), which assigns the same priors to all agents.[15] Since ideally rational agents' credences are only functions of priors and evidence (Teller 1973; Armendt 1980), in the absence of Byzantine Failures inconsistencies among agents would all be due to different evidence. Then Conditionalization and Reflection have no trouble meeting Christensen (2000)'s demand for impartiality, and no Double Agent Dutch Books are possible against such strictly rational agents.

This is an extremely strict model for ideal rationality. Philosophers who want to take ideal rationality as normative for humans may naturally rebel at such a model.[16] But humans are subject to Byzantine Faults. A model of bounded rationality intended to be normative for humans must show how those faults can be prevented from developing into Byzantine Failures. This will inevitably mean deciding that in certain situations insufficient parallelism is available for any claim to consistency. In other words, there will be situations in which agents with such bounded rationality will not bet. It is irrational to visit a bookie with your partner if you think you have opposing beliefs and a joint checking account, and it is just as irrational to bet when you suspect that you are experiencing a psychological difficulty that impedes your rationality. Nor should we have expected human-like agents to accept bets at some fair betting quotient on all propositions, since human-like agents obviously lack the complete set of probabilistic beliefs necessary to have such quotients. The characteristics of ideal rational agents are closely intertwined, and rejecting some of those characteristics on the strength of arbitrary value faults without considering what the possibility of such faults says about the system as a whole only leads to confusion.

[15] Necessary a priori priors combined with diachronic Dutch strategies and Conditionalization also solves the Sleeping Beauty problem (Milano 2022), another point in favor of such a demanding standard.

[16] Though as John Broome (2007) points out, it can be quite difficult to justify such desires.

4. References

Armendt, Brad. 1980. “Is There a Dutch Book Argument for Probability Kinematics?” *Philosophy of Science* 47 (4): 583–88. https://doi.org/10.1086/288958.

Arora, A., and S. S. Kulkarni. 1998. “Detectors and Correctors: A Theory of Fault-Tolerance Components.” In *Proceedings of the 18th International Conference on Distributed Computing Systems*, 436–43. https://doi.org/10.1109/ICDCS.1998.679772.

Arrow, Kenneth J. 1950. “A Difficulty in the Concept of Social Welfare.” *Journal of Political Economy* 58 (4): 328–46.

Backes, Marvin. 2019. “A Bitter Pill for Closure.” *Synthese* 196 (9): 3773–87. https://doi.org/10.1007/s11229-017-1620-8.

Biely, Martin, and Martin Hutle. 2009. “Consensus When All Processes May Be Byzantine for Some Time.” In *Stabilization, Safety, and Security of Distributed Systems*, edited by Rachid Guerraoui and Franck Petit, 120–32. Lecture Notes in Computer Science. Berlin: Springer.

Briggs, R. 2010. “Decision-Theoretic Paradoxes as Voting Paradoxes.” *The Philosophical Review* 119 (1): 1–30.

Broome, John. 2007. “Is Rationality Normative?” *Disputatio* 2 (23): 161–78. https://doi.org/10.2478/disp-2007-0008.

Christensen, David. 1991. “Clever Bookies and Coherent Beliefs.” *The Philosophical Review* 100 (2): 229–47. https://doi.org/10.2307/2185301.

———. 2000. “Diachronic Coherence versus Epistemic Impartiality.” *The Philosophical Review* 109 (3): 349–71. https://doi.org/10.2307/2693694.

———. 2007a. “Epistemic Self-Respect.” *Proceedings of the Aristotelian Society* 107 (1pt3): 319–37. https://doi.org/10.1111/j.1467-9264.2007.00224.x.

———. 2007b. “Epistemology of Disagreement: The Good News.” *The Philosophical Review* 116 (2): 187–217.

Driscoll, K., B. Hall, M. Paulitsch, P. Zumsteg, and H. Sivencrona. 2004. “The Real Byzantine Generals.” In *Proceedings of the 23rd Digital Avionics Systems Conference*, 2:6.D.4-61. https://doi.org/10.1109/DASC.2004.1390734.

Fama, Eugene F. 1965. “Random Walks in Stock Market Prices.” *Financial Analysts Journal* 21 (5): 55–59.

Finetti, B. de. 1964. “Foresight: Its Logical Laws, Its Subjective Sources.” In *Studies in Subjective Probability*, edited by Henry E. Kyburg, Jr. and Howard E. Smokler, 1st edition. John Wiley and Sons.

Fraassen, C. van. 1984. “Belief and the Will.” *The Journal of Philosophy* 81 (5): 235–56. https://doi.org/10.2307/2026388.

Genot, Emmanuel J., and Justine Jacot. 2020. “The Brain Attics: The Strategic Role of Memory in Single and Multi-Agent Inquiry.” *Synthese* 197 (3): 1203–24. https://doi.org/10.1007/s11229-018-1743-6.

Grüne-Yanoff, Till. 2007. “Bounded Rationality.” *Philosophy Compass* 2 (3): 534–63. https://doi.org/10.1111/j.1747-9991.2007.00074.x.

Howson, Colin. 1992. “Dutch Book Arguments and Consistency.” *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1992 (2): 161–68. https://doi.org/10.1086/psaprocbienmeetp.1992.2.192832.

Kuznets, Roman, Laurent Prosperi, Ulrich Schmid, and Krisztina Fruzsa. 2019. “Epistemic Reasoning with Byzantine-Faulty Agents.” In *Frontiers of Combining Systems*, edited by Andreas Herzig and Andrei Popescu, 259–76. Lecture Notes in Computer Science.

Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-29007-8_15.
Lamport, Leslie, Robert Shostak, and Marshall Pease. 1982. “The Byzantine Generals Problem.” *ACM Transactions on Programming Languages and Systems* 4 (3): 382–401. https://doi.org/10.1145/357172.357176.
Lang, Patrick. 2014. “Bayesian Epistemology of Disagreement.” M.A. Thesis, Vienna, Austria: University of Vienna. http://othes.univie.ac.at/31638/1/2014-01-15_0749501.pdf.
Lewis, David. 1980. “A Subjectivist’s Guide to Objective Chance.” In *Studies in Inductive Logic and Probability, Volume II*, edited by Richard C. Jeffrey, 263–93. Berkeley: University of California Press.
May, Kenneth O. 1952. “A Set of Independent Necessary and Sufficient Conditions for Simple Majority Decision.” *Econometrica* 20 (4): 680–84. https://doi.org/10.2307/1907651.
Milano, Silvia. 2022. “Bayesian Beauty.” *Erkenntnis* 87 (2): 657–76. https://doi.org/10.1007/s10670-019-00212-4.
Milne, Peter. 1991. “A Dilemma for Subjective Bayesians? And How to Resolve It.” *Philosophical Studies* 62 (3): 307–14. https://doi.org/10.1007/BF00372396.
Neumann, John von, and Oskar Morgenstern. 1953. *Theory of Games and Economic Behavior*. 3rd Edition. Princeton University Press. https://www.jstor.org/stable/j.ctt1r2gkx.
Pease, M., R. Shostak, and L. Lamport. 1980. “Reaching Agreement in the Presence of Faults.” *Journal of the Association for Computing Machinery* 27 (2): 228–34. https://doi.org/10.1145/322186.322188.
Pettigrew, Richard. 2016. “Accuracy, Risk, and the Principle of Indifference.” *Philosophy and Phenomenological Research* 92 (1): 35–59. https://doi.org/10.1111/phpr.12097.
Ramsey, F.P. 1964. “Truth and Probability.” In *Studies in Subjective Probability*, edited by Henry E. Kyburg, Jr. and Howard E. Smokler, 1st edition. John Wiley and Sons.
Roush, Sherrilyn. 2009. “Second Guessing: A Self-Help Manual.” *Episteme* 6 (3): 251–68. https://doi.org/10.3366/E1742360009000690.
Ryder, J. M. 1981. “Consequences of a Simple Extension of the Dutch Book Argument.” *The British Journal for the Philosophy of Science* 32 (2): 164–67. https://doi.org/10.1093/bjps/32.2.164.
Skyrms, Brian. 1987. “Coherence.” In *Scientific Inquiry in Philosophical Perspective*, edited by Nicholas Rescher, 225–42. Pittsburgh, PA: University of Pittsburgh Press. http://fitelson.org/probability/skyrms_coherence.pdf.
———. 1993. “A Mistake in Dynamic Coherence Arguments?” *Philosophy of Science* 60 (2): 320–28.
Sobel, Jordan Howard. 1987. “Self-Doubts and Dutch Strategies.” *Australasian Journal of Philosophy* 65 (1): 56–81. https://doi.org/10.1080/00048408712342771.
Talbott, W. J. 1991. “Two Principles of Bayesian Epistemology.” *Philosophical Studies* 62 (2): 135–50.
Teller, Paul. 1973. “Conditionalization and Observation.” *Synthese* 26 (2): 218–58. https://doi.org/10.1007/BF00873264.
Vineberg, Susan. 1997. “Dutch Books, Dutch Strategies and What They Show about Rationality.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 86 (2): 185–201.
———. 2016. “Dutch Book Arguments.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2016/entries/dutch-book/.

Weatherson, Brian. 2004. “Some Dutch Book Arguments.” *Thoughts Arguments and Rants* (blog). September 30, 2004. http://tar.weatherson.org/2004/09/30/some-dutch-book-arguments/.

———. 2005. “What Do Dutch Book Arguments Prove.” *Thoughts Arguments and Rants* (blog). July 10, 2005. http://tar.weatherson.org/2005/07/10/what-do-dutch-book-arguments-prove/.

Williams, Daniel. 2021. “Epistemic Irrationality in the Bayesian Brain.” *The British Journal for the Philosophy of Science* 72 (4): 913–38. https://doi.org/10.1093/bjps/axz044.

Wolfers, Justin, and Eric Zitzewitz. 2006. “Interpreting Prediction Market Prices as Probabilities.” National Bureau of Economic Research Working Paper 12200. https://doi.org/10.3386/w12200.