Experimental Demonstration of Quantum Effects in the Operation of Microscopic Heat Engines

James Klatzow,1 Jonas N. Becker,1 Patrick M. Ledingham,1 Christian Weinzel,1 Krzysztof T. Kaczmarek,1,2 Dylan J. Saunders,1 Joshua Nunn,3 Ian A. Walmsley,1,* Raam Uzdin,4 and Eilon Poem5,†
1Clarendon Laboratory, University of Oxford, Parks Road, Oxford OX1 3PU, United Kingdom
2Groupe de Physique Appliquée, Université de Genève, CH-1211 Genève, Switzerland
3Centre for Photonics and Photonic Materials, Department of Physics, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom
4Fritz Haber Research Center for Molecular Dynamics, Hebrew University of Jerusalem, Jerusalem 9190401, Israel
5Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 7610001, Israel

(Received 3 October 2018; revised manuscript received 7 January 2019; published 20 March 2019)

The ability of the internal states of a working fluid to be in a coherent superposition is one of the basic properties of a quantum heat engine. It was recently predicted that in the regime of small engine action, this ability can enable a quantum heat engine to produce more power than any equivalent classical heat engine. It was also predicted that in the same regime, the presence of such internal coherence causes different types of quantum heat engines to become thermodynamically equivalent. Here, we use an ensemble of nitrogen vacancy centers in diamond for implementing two types of quantum heat engines, and experimentally observe both effects.

DOI: 10.1103/PhysRevLett.122.110601

A heat engine uses a working fluid connected to hot and cold reservoirs to generate power. In a quantum heat engine, the working fluid has some properties that cannot be described by classical physics. The most basic of these is the ability of internal energy states of the working fluid to be in a coherent superposition.

This concept of a quantum-coherent heat engine was introduced already sixty years ago by Scovil and Schulz-DuBois [1], who linked the efficiency of a three-level maser to the Carnot efficiency.

One question that then arises is whether the presence of coherence between two internal energy states of an engine can enhance its performance.

At first, it looked like the answer to this question was negative: it was proven that internal coherence cannot lead to efficiencies higher than the Carnot limit [2–4], and the first coherence-related effect predicted for heat engines, the “quantum friction,” actually involved degradation in performance [5–10]. Furthermore, when performance enhancement was first predicted, it was not due to internal quantum properties of the system, but rather to the use of nonthermal energy sources such as externally injected coherence [11–13] or squeezed baths [14–16].

More recently, however, Uzdin and co-workers [17] have theoretically shown that internal coherent superposition states created during the operation of a quantum heat engine affect measurable thermodynamic quantities, and that their presence can be performance enhancing on its own—that is, when using only standard thermal baths.

Microscopic heat machines have been recently implemented with trapped ions [18,19], and proposals for heat machines using superconducting circuits [20,21] and optomechanics [22,23] have been made. However, when operated with thermal baths, the machines implemented so far have not demonstrated any inherently quantum feature in their thermodynamic quantities.

Here we measure such features in microscopic heat engines. We use an ensemble of negatively charged nitrogen vacancy (NV−) centers in diamond to implement two types of quantum heat engines, a two-stroke engine and a continuous engine, and experimentally observe the two coherent superposition-related effects predicted in Ref. [17]: (i) a quantum thermodynamic signature (QTS)—an increased output power of a quantum engine with respect to that of any classical engine using the same resources, and (ii) quantum heat-machine equivalence (QHME)—a convergence in the output power of the two different quantum heat engine types.

In general, the cycle of a quantum heat engine consists of a sequence of operations (strokes), which include the interaction of the system either with a thermal bath, or with an external field acting as the work reservoir. While interactions with the thermal baths may only change the populations of the energy states of the heat engine without generating coherence, the unitary evolution due to the field may bring the energy states into a coherent superposition. This internal generation of coherence is part of the dynamics induced by the driving field, and unlike external coherence injection [11–13], preserves the entropy of the system [17].

The change in the state of the system during a stroke can be quantified by the action of the stroke, which is defined to
be the time integral over the stroke of the norm of the generator of motion [17]. For constant-coupling strokes, the stroke action, $s_i$, is simply given (in units of $\hbar$) by $s_i = \gamma_i \tau_i$, where $\tau_i$ is the stroke duration, and $\gamma_i$ is the total coupling rate of the system to the bath(s) (i.e., the total population transfer rate) and/or to the driving field (i.e., the Rabi frequency) during the stroke. The total action per cycle, $s$, is then given by $s = \sum_i s_i$.

By changing the scheduling and durations of the strokes, it is possible to implement different engine varieties. Figure 1(a) schematically presents three basic engine types for a three-level system. The box with the letter $U$ represents the unitary operation applied during the work stroke, and the red (blue) arrow represents population transfer due to coupling to a hot (cold) bath. The continuous engine (left), where all operations are performed together in a continuous fashion, most resembles the continuous-wave maser discussed by Scovil and Schulz-DuBois. In the four-stroke cycle (right), the couplings to the hot bath, cold bath, and the work repository are completely separated. As shown in the figure, this engine coincides with the classical Otto engine when the action of $U$ is a full population swap [17]. The two-stroke engine (center) is an intermediate case where the coupling to the work repository is separated from the couplings to the baths, but the system is coupled to both baths together. See Sec. S1 of Ref. [24] for more details.

In the small-action regime, $s \ll 1$, i.e., when the stroke durations are much shorter than the typical coupling times of the system to the thermal baths and the work repository, two coherence-related phenomena occur [17].

First, the output power is larger when coherence is present at the beginning of the work strokes—that is, when the coherent superposition does not completely dephase during the coupling to the baths. This effect constitutes a direct signature of coherence in a thermodynamic quantity—a QTS. It stems from the fact that the rate of change of population by the external field, required for the production of work, is proportional to the existing coherence [35]. In the small-action regime, where the external field is weak enough such that significant coherence cannot be built within a single engine cycle, but only over many cycles, the work per cycle, $W_{\text{cyc}}$, of a coherent engine, where coherence is present at the beginning of the cycle, will be linear in the cycle time, $\tau_{\text{cyc}}$, while that of a stochastic engine, where there is no initial coherence and coherence has to be built up during the cycle, will be quadratic in $\tau_{\text{cyc}}$ [17]. This means that in this regime, the average power output, $\langle P \rangle = W_{\text{cyc}}/\tau_{\text{cyc}}$, will be constant for coherent engines, but go down linearly with $\tau_{\text{cyc}}$ for stochastic engines. For the case of a two-stroke engine, it can be shown (see Ref. [17] or Sec. S8 of Ref. [24]) that the output power of any stochastic engine has to obey the following bound: $\langle P \rangle_{\text{stoch}} \leq \frac{1}{2} \hbar \omega_{10} d^2 \Omega^2 \tau_{\text{cyc}}$. Here $\omega_{10}$ is the angular frequency of the transition between the two levels interacting with the external field, $\Omega$ is the Rabi frequency, and $d$ is the ratio between the work stroke duration and $\tau_{\text{cyc}}$ (the duty cycle). For a short enough $\tau_{\text{cyc}}$, the power of a coherent engine will surpass this stochastic bound.

Second, the work and power of coherent engines in the small-action regime do not depend on the type of the engine: all coherent engines using the same resources would be thermodynamically equivalent. This QMHE effect is due to the nonvanishing coherence in the energy basis, combined with the time-symmetric structure of the stroke protocols. When expanding the expression for the work per work stroke of a coherent engine in powers of the action of that work stroke, $s_{wi}$, the second-order contribution vanishes due to symmetry, and the first-order contribution is directly proportional to the coherence at the beginning of the work stroke, $C_0$, which, for $s \ll 1$, is the same for all work strokes in the cycle (see Ref. [17] and Sec. S9 of Ref. [24]). That is, $W_{\text{cyc}} \propto C_0 s_w + O(s_w^3)$, where $s_w = \sum_i s_{wi}$. Therefore, all coherent engines using the same resources, and thus having the same $s_w$ and $C_0$, will yield the same $W_{\text{cyc}}$, regardless of the engine type. In contrast, for stochastic engines, as $C_0$ is replaced by the coherence generated within a work stroke, which is of $O(s_{wi})$, one obtains $W_{\text{cyc}} \propto \sum_i s_{wi}^2$. This nonlinear dependence can lead to different work outputs for different engine types. For example, a stochastic two-stroke engine with a work-stroke time $\tau_w$ would yield $W_{\text{cyc}} \propto \tau_w^2$, while its equivalent stochastic four-stroke engine, having two ($\tau_w/2$)-long work strokes, would yield a lower work of $W_{\text{cyc}} \propto 2(\tau_w/2)^2 = \tau_w^2/2$ [17].

For demonstrating the quantum heat engine effects discussed above, we use the diamond NV$^-$ center [36], which is an atomic-like system that exhibits several features that make it suitable for this purpose.

First, its ground-state manifold contains three spin states, $| -1 \rangle$, $| 0 \rangle$, and $| +1 \rangle$, that can maintain coherence for a substantial amount of time even at room temperature, and can coherently interact with a microwave (MW) field serving as the work repository (the "load").

Second, after optical excitation, the system decays back into the ground-state manifold both by spin-preserving radiative decay, and by spin-nonpreserving nonradiative channels through a metastable spin-singlet state $| 0' \rangle$ [Fig. 1(b)]. The system therefore tends to a steady state with a population difference between the different ground-state spin components. It can be shown (see Sec. S6 of Ref. [24]) that the dynamics of this process is equivalent to that produced by coupling to Markovian heat baths, and the effective system-environment coupling rates can be calculated for the relevant optical excitation rates (see Sec. S6 of Ref. [24]). Thus, this system qualifies for demonstrating the above mentioned effects, relevant for any Markovian environment.

Third, due to the spin dependence of the nonradiative decay channels, the fluorescence intensity provides a direct means to measure the populations within the ground-state
manifold. This technique, known as optically detected magnetic resonance [36], provides better sensitivity than direct measurement of microwave amplification, allowing us to work in the small-action regime.

As shown in Fig. 1(b) (left), a fixed, external magnetic field of 0.2 T, applied along the NV– symmetry axis, lowers the energy of the state \( |+1 \rangle \) below that of \( |0 \rangle \) by the Zeeman interaction. Due to spin-nonpreserving decay, the effective rate of population excitation from \( |0 \rangle \) to \( |0' \rangle \) is much lower than the decay rate from \( |0' \rangle \) to \( |0 \rangle \), while the effective excitation rate from \( |+1 \rangle \) to \( |0' \rangle \) is almost equal to its corresponding decay rate [37]. Thus, when the light is on, the \( |0 \rangle \rightarrow |0' \rangle (|+1 \rangle \rightarrow |0' \rangle) \) subsystem can be seen as coupled to a cold (hot) bath (see Sec. S6 of Ref. [24]). This results in population inverison between \( |+1 \rangle \) and \( |0 \rangle \) [Fig. 1(b), right]. An external MW field, resonantly exciting only the \( |+1 \rangle \leftrightarrow |0 \rangle \) transition, can then extract work, in the form of stimulated emission of MW radiation. Note that as all MW transitions involving the state \( |−1 \rangle \) are out of resonance, its contribution to the work can be neglected, and the effective system contains only three levels: \( |+1 \rangle, |0 \rangle, \) and \( |0' \rangle \).

The optical excitation inducing the effective thermal interaction and the MW driving manifesting the coupling to the work repository can either be interlaced or both be on continuously, implementing either a two-stroke or a continuous engine, respectively.

The experimental system consisted of a single crystal diamond sample containing a dense \((\sim 10^{18}\text{ cm}^{-3})\) ensemble of NV centers, two permanent magnets that applied a constant magnetic field along the [111] crystal direction, a MW generator and a MW strip-line waveguide for MW excitation, a green laser for optical excitation, and a confocal geometry light collection and detection setup. An acousto-optic modulator (AOM) was used for the intensity modulation of the laser light reaching the sample, and a fast MW switch (MS) was used for the amplitude modulation of the applied MWs. The MS and the AOM had switching times of 1.5 ns and 12 ns, respectively, and were both simultaneously driven at rates of a few MHz by a fast function generator. To detect the change in fluorescence due to the operation of the engine—that is, due to the MW driving—lock-in detection was used. The output of the MW generator was amplitude modulated by a \( \sim 100\) Hz square wave, much slower than the repetition rates of the engine, and the photodiode signal at this sideband was registered. More details on the sample, setup, and setup calibration are given, respectively, in Secs. S2, S3, and S4 of Ref. [24].

As the only ways the MW field can change level population and affect the fluorescence rate are through absorption or stimulated emission, the measured change in fluorescence could be converted into average output engine power, in terms of the number of emitted MW photons per second per NV– center. This was achieved by using the calibrated excitation rate, the known decay rates of the system, and a rate equation model. See Secs. S5, S7, and S10 of Ref. [24] for more details.

These measurement techniques, in combination with the use of a dense NV– ensemble, allowed us to work with the very low actions required for demonstrating QTS and QHME, a regime that has not been explored in previous NV– experiments [36].

Figure 2(a) presents the measured power output of the two-stroke engine vs the action per cycle (varied by changing the cycle time between 30 ns and 180 ns), for \( \Omega = 2\pi \times (0.25 \pm 0.01)\text{ MHz}, d = 1/3, \) and a total thermal coupling rate of \( \gamma_{\text{th}} = 1.76 \pm 0.08\text{ MHz} \) (see Sec. S6 of Ref. [24]), along with the relevant stochastic bound.

For the smallest action applied (dashed frame, enlarged in the inset), the bound is violated by 2.4 standard deviations, corresponding to a single-sided \( p \) value of 0.0082 (see Sec. S11 of Ref. [24]). This is a clear indication of a QTS.

We also studied the work output as the coherence of the system is reduced. Due to inhomogeneous broadening of the energy difference between the states \( |0 \rangle \) and \( |+1 \rangle \), the engine experiences pure dephasing on a timescale of \( T_2^* \sim 75\text{ ns} \) (see Sec. S4 of Ref. [24]). Thus, by changing the length of the thermal stroke, the amount of pure dephasing during that stroke can be controlled. To ensure that the amount of pure dephasing is the only parameter that is varied, the optical
excitation rate is lowered as the length of the thermal stroke is increased, such that the population transfer action of the thermal stroke is kept constant. The work stroke is also fixed, with a duration of 10 ns and a Rabi frequency of $2\pi \times 0.25$ MHz. The total action per cycle excluding pure dephasing is thus fixed, at a value of $0.05\hbar$.

Figure 2(b) presents the work per cycle in the two-stroke engine (red) as a function of the thermal stroke duration in units of dephasing time $T_2^* = 75$ ns. Red dots (line) present the measured data (theoretical prediction). The blue line is the stochastic bound. The work-stroke length and Rabi frequency are fixed, while the optical excitation rate is adjusted to keep the thermal action constant. The insets on the left (right) depict cycles with a short (long) thermal stroke. The measured work output decreases to below the bound.

Excitation rate is lowered as the length of the thermal stroke is increased, such that the population transfer action of the thermal stroke is kept constant. The work stroke is also fixed, with a duration of 10 ns and a Rabi frequency of $2\pi \times 0.25$ MHz. The total action per cycle excluding pure dephasing is thus fixed, at a value of $0.05\hbar$.

Figure 2(b) presents the work per cycle in the two-stroke engine (red) as a function of the thermal stroke duration in units of dephasing time $T_2^* = 75$ ns. Red dots (line) present the measured data (theoretical prediction). The blue line is the stochastic bound. The work-stroke length and Rabi frequency are fixed, while the optical excitation rate is adjusted to keep the thermal action constant. The insets on the left (right) depict cycles with a short (long) thermal stroke. The measured work output decreases to below the bound.

It is clearly seen that the output work per cycle decreases as the thermal stroke duration approaches the dephasing time, and it drops below the stochastic bound for sufficiently long strokes. This bound increases slightly at long thermal strokes, taking into account experimental imperfections. The slight discrepancy between theory and measurement for long thermal strokes might be attributed either to homogeneous dephasing or to charging effects, both neglected in the present theory (See Sec. S7 of Ref. [24]).

These measurements clearly show that in the small-action regime, the presence of coherence in microscopic heat engines is manifested in their output work and power, allowing them to operate more efficiently than any corresponding classical engine.

Figure 3 presents a comparison between the power outputs of two-stroke and continuous engines. The symbols present the power outputs of a two-stroke engine vs its action per cycle, varied by changing its cycle time, for $\gamma_{th} = 1.76 \pm 0.08$ MHz and several values of $\Omega$, as indicated. The duty cycle was fixed at $d = 1/3$. For each data set, the power output of a continuous engine with the same mean Rabi frequency and mean effective thermalization rate as in the corresponding two-stroke engine is presented by the horizontal, full rectangles. The widths of the rectangles signify the measurement error. The clear convergence in the performance of the two engine types for small two-stroke engine actions constitute the first experimental verification of QHME. The theoretical predictions for the output power of the two-stroke (continuous) engine are presented in Fig. 3 by the solid (dashed) lines (see Sec. S7 of Ref. [24]).

In conclusion, we have used an ensemble of NV$^-$ centers in diamond for implementing two different types of
quantum heat engines, and we have observed both a QTS and the QHME. This constitutes the first experimental measurement of quantum effects in heat machines. While the theory in Ref. [17], derived for an idealized three-level system with ideal Markovian environments, can be verified analytically, this work demonstrates that a physical system with many levels can satisfy all the assumptions needed in the theoretical derivation, and it shows that these effects are experimentally accessible. We therefore hope that this work will motivate further research along at least three lines. (i) Demonstration of quantum effects in other physical realizations of heat machines such as superconducting circuits [20,21] and ion traps [18,19]. Hopefully, with these it will be possible to verify the validity of the QHME also for transient dynamics [17] and for four-stroke machines. (ii) Theoretical search for quantum thermodynamic signatures in heat machines based on other quantum characteristics such as entanglement [38], the violation of Leggett-Garg inequalities [39,40], and quantum discord. (iii) Application to the design and development of novel devices such as room-temperature masers [41,42]. We further hope that this work will be of interest to other research areas concerned with the role of quantum coherence in the enhancement of work extraction by microscopic heat engines, such as the study of photosynthesis [43] and the development of solar cells.

The authors would like to thank Ran Fischer and Paz London for their assistance with sample preparation. This work was supported by the United Kingdom EPSRC (No. EP/I000051/1), the U.S. AFOSR EOARD (No. FA8655-09-1-3020), EU IP SIQS (No. 600645), EU COST (Action No. MP1209), and an EU Marie Curie Fellowship (No. IEF-2013-627372 to E.P.). P.M. L. acknowledges a European Union Horizon 2020 Research and Innovation Framework Programme Marie Curie individual fellowship, Grant Agreement No. 705278. J.N. acknowledges support from a Royal Society University Research Fellowship. I.A.W. acknowledges an ERC Advanced Grant (MOQUACINO) and an EPSRC Programme Grant (No. EP/K034480/1). J.N.B. was supported by the EPSRC Networked Quantum Information Technologies Hub (NQIT).

1ian.walmsley@physics.ox.ac.uk
2eilon.poem@weizmann.ac.il


